Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models

Cheryl J. Flynn, Clifford M. Hurvich, and Jeffrey S. Simonoff

New York University

February 11, 2013

Abstract

It has been shown that AIC-type criteria are asymptotically efficient selectors of the tuning parameter in non-concave penalized regression methods under the assumption that the population variance is known or that a consistent estimator is available. We relax this assumption to prove that AIC itself is asymptotically efficient and we study its performance in finite samples. In classical regression, it is known that AIC tends to select overly complex models when the dimension of the maximum candidate model is large relative to the sample size. Simulation studies suggest that AIC suffers from the same shortcomings when used in penalized regression. We therefore propose the use of the classical corrected AIC (AIC_c) as an alternative and prove that it maintains the desired asymptotic properties. To broaden our results, we further prove the efficiency of AIC for penalized likelihood methods in the context of generalized linear models with no dispersion parameter. Similar results exist in the literature but only for a restricted set of candidate models. By employing results from the classical literature on maximum-likelihood estimation in misspecified models, we are able to establish this result for a general set of candidate models. We use simulations to assess the performance of AIC and AIC_c, as well as that of other selectors, in finite samples for both SCAD-penalized and Lasso regressions and a real data example is considered.

KEY WORDS: Akaike information criterion; Least absolute shrinkage and selection operator (Lasso); Model selection/ Variable Selection; Penalized likelihood; Smoothly clipped absolute deviation (SCAD).

1 Introduction

Regularized (or penalized) likelihood methods have become widely used in recent years due to the increased availability of large data sets. These methods operate by maximizing the penalized likelihood function

$$\frac{1}{n}l(\boldsymbol{\beta}) - \sum_{j=1}^{d_n} p_\lambda(|\beta_j|) \tag{1.1}$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^{d_n}$, where $l(\beta)$ is the working log-likelihood function, d_n is the total number of predictors, and $p_\lambda(\cdot)$ is a penalty function that penalizes against model complexity and the size of the estimated coefficients. The working log-likelihood is used to justify the first part of the function (e.g., in Least Squares, the working log-likelihood is based on the Gaussian distribution). As demonstrated in Sections 2 and 3, many of the results discussed in this paper are valid even if the working log-likelihood is misspecified. With these methods, increasing the amount of regularization increases the number of estimated coefficients that are set equal to zero thus performing "automatic" variable selection through the data-dependent choice of the regularization parameter, λ. In contrast, variable selection in classical regression is commonly done using the Leaps and Bounds algorithm (Furnival and Wilson, 1974), which becomes infeasible when the number of predictors is much larger than 30 (Hastie et al., 2009). For most penalty functions efficient algorithms exist to compute the estimated models over a regularization path making it possible to do variable selection in high dimensions.

The performance of the estimated model heavily depends on the choice of the regularization parameter. In regularized regression several classical model selection procedures have been heuristically applied as selectors of this parameter including information criteria such as Akaike's information criterion (AIC; Akaike, 1973), the Bayesian information criterion (BIC; Schwarz, 1978), and Generalized cross-validation (GCV; Craven and Wahba, 1978) as well as data-based selection procedures such as k-fold cross-validation (see, e.g., Fan and Li, 2001, Zou et al., 2007, Wang et al., 2007, and Zhang et al., 2010 for applications of these selectors to penalized regression estimators). The statistical properties of these model selection procedures have been widely studied in the context of classical regression and an ongoing research problem is to determine if these properties carry over to the context of penalized regression.

The asymptotic performance of model selection procedures can be studied under two important and distinct settings: (1) when the true model is not among the candidate models (the "non-true model world") and (2) when the true model is among the candidate models (the "true model world"). In the non-true model world a reasonable goal is *efficient* model selection, meaning that we would like to select the model that asymptotically performs the best amongst the candidate models. In contrast, in the true-model world most of the literature focuses on *consistent* model selection, meaning that the probability that the true model is chosen is asymptotically one. In general, a model selection procedure cannot be both consistent and efficient (Shao, 1997; Yang, 2005). Although the non-true model world has been extensively studied in classical regression (e.g., Shibata, 1981, Li, 1987, Hurvich and Tsai, 1989, 1991, Shao, 1997, and Burnham and Anderson, 2002) the majority of the research on model selection in penalized regression has focused on the true model world (e.g., Leng et al., 2006, Zou et al., 2007, and Wang et al., 2007). We feel that the non-true model world is more realistic in many situations since the data-generating process is likely to be too complex to know exactly; this is the essence of George Box's famous admonition that "all models are wrong, but some are useful" (Box, 1979). This setting should be of particular interest to researchers and data analysts in areas such as social science and environmental health where a large number of predictors are expected to influence the dependent variable (too many to include in model fitting; Gelman, 2010) as well as machine learning where the goal is typically not to uncover the true data generating process but rather to find a model that can predict well.

In the context of generalized linear models (GLMs), Zhang et al. (2010) (hereafter ZLT) proposed the use of a "GIC-type" criterion,

$$GIC_{\kappa_n} = -\frac{1}{n} l(\hat{\boldsymbol{\beta}}_\lambda) + \kappa_n \frac{df_\lambda}{n}$$

for choosing the regularization parameter λ for non-concave penalized estimators in both the non-true model world and the true-model world. Here $\hat{\boldsymbol{\beta}}_\lambda$ is the estimator that maximizes (1.1) for a specific λ, df_λ is the effective degrees of freedom and the log-likelihood function corresponds to a member of the exponential family, i.e.

$$l(\hat{\boldsymbol{\beta}}_\lambda) = \sum_{i=1}^{n} \left(\frac{y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda - b(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda)}{a(\phi)} + c(y_i, \phi) \right),$$

where the form of functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ depends on the specified distribution and ϕ is the dispersion parameter (see e.g. McCullagh and Nelder, 1989). They showed that "AIC-type" versions of GIC_{κ_n} ($\kappa_n \to 2$) are efficient in the former case, while "BIC-type" versions of GIC_{κ_n} ($\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$) are consistent in the latter case.

In the Gaussian model, GIC_{κ_n} takes on a form that includes the true error variance σ^2, and the proofs operate under the assumption that this is known or that a consistent estimator is available. However, if the true model is not included in the set of candidate models then a consistent estimator of the true error variance may not be available (Shao, 1997) making the efficiency proofs of ZLT not applicable in practice. This motivates us to extend the ZLT results in various ways. First, we show that the feasible version of GIC_2, which corresponds to the well-known C_p measure (Mallows, 1973), is in fact efficient in the non-true model world. Second, we show that AIC and GCV, which do not require a consistent estimator of σ^2, are also efficient. Third, we show that although several model selection procedures may be asymptotically optimal, performance varies in finite samples. Specifically, we study performance when the number of predictors is allowed to be large relative to the sample size and show that AIC, BIC, C_p, and GCV all have a tendency to sometimes catastrophically overfit (lead to λ values approaching 0). In classical regression Hurvich and Tsai (1989) showed that AIC has a tendency to select overly complex models when the dimension of the maximum candidate model is large relative to the sample size and proposed a corrected version of AIC (AIC_c). We show that AIC_c is also efficient, but avoids the tendency to select overly complex models. We use Monte Carlo simulations to illustrate the properties of these methods in finite samples and compare their performance against the data-dependent method 10-fold CV.

For GLMs where there is no dispersion parameter (e.g., probit and logistic regression or the Poisson log-linear model), there is no difference between GIC_2 and AIC. However, in their proof ZLT restrict the set of candidate models to ones where the estimated parameter converges in probability to the true parameter uniformly. To weaken this assumption we employ the result from White (1982) that the maximum-likelihood estimator converges almost surely to a "pseudo-true" parameter (the parameter that minimizes the Kullback-Leibler (KL) loss function) when the model is misspecified and prove the efficiency of AIC under a weaker set of assumptions. These results, and the results for the Gaussian model, apply to a wide range of penalized likelihood estimators, including both non-concave penalized estimators and the well-known Least absolute shrinkage and selection operator (Lasso) estimator (Tibshirani, 1996).

The remainder of the paper is organized as follows. Section 2 focuses on penalized regression and establishes the efficiency results for C_p, AIC, GCV and AIC_c without the assumption that the true population variance is known or that a consistent estimator exists. Section 3 focuses on GLMs where there is no dispersion parameter and establishes the efficiency of AIC for a general set of candidate models. Section 4 presents simulation results that explore the finite-sample behavior of the different selectors when the number of predictors is allowed to be large relative to the sample size. An empirical example that highlights the varying performance of the selectors is presented in Section 5. Concluding remarks are given in Section 6. The main proofs are included in the appendix with some auxiliary results included in the supplementary material.

2 Gaussian Model

For ease of notation, in this section, and for the remainder of the paper, we suppress the subscript n where we feel it is clear that a variable depends on the sample size.

To study model selection in regularized regression we consider the model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the $n \times 1$ response vector, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is a $n \times 1$ unknown mean vector and the entries of the $n \times 1$ error vector $\boldsymbol{\varepsilon}$ are independent and identically distributed (iid) with mean 0 and variance σ^2. The mean vector is estimated by $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$ where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is a $n \times d_n$ deterministic matrix of predictors and $\hat{\boldsymbol{\beta}}_\lambda$ is the estimator that minimizes the penalized least squares function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \sum_{j=1}^{d_n} p_\lambda(|\beta_j|)$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^{d_n}$.

Adopting the notation from ZLT, we let the index set $\mathcal{A}_n$ denote the class of all candidate models and we assume that $\bar{\alpha} = \{1, \ldots, d_n\}$ is the largest model in $\mathcal{A}_n$. For any $\alpha \in \mathcal{A}_n$, we define d_α to be the number of predictor variables included in the candidate model. We further define the least squares estimated mean vector by $\hat{\boldsymbol{\mu}}_\alpha = \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha$ where $\mathbf{X}_\alpha$ is the matrix of predictors that are included in candidate model α and $\hat{\boldsymbol{\beta}}_\alpha$ is the corresponding vector of the estimated least squares coefficients. The associated projection matrix is $\mathbf{H}_\alpha = \mathbf{X}_\alpha(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha$. For a given λ, we define α_λ to be the model $\alpha \in \mathcal{A}_n$ whose predictors are those with non-zero coefficients in the penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$ and let df_λ denote the effective degrees of freedom. The least squares estimated mean vector based on the model α_λ is denoted by $\hat{\boldsymbol{\mu}}_{\alpha_\lambda} = \mathbf{X}_{\alpha_\lambda}\hat{\boldsymbol{\beta}}_{\alpha_\lambda}$. In this equation, $\mathbf{X}_{\alpha_\lambda}$ is the matrix of predictors whose coefficients are not shrunk to zero in the penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}$ are the estimated coefficients from the least squares model fit using these predictors. The associated projection matrix in this case is defined as $\mathbf{H}_{\alpha_\lambda} = \mathbf{X}_{\alpha_\lambda}(\mathbf{X}'_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda})^{-1}\mathbf{X}'_{\alpha_\lambda}$.

If we assume that we are in the non-true model world, then a reasonable goal is efficient model selection. The L_2 loss is commonly used to assess the predictive performance of an estimator and is calculated as

$$L(\hat{\boldsymbol{\beta}}_\lambda) = \frac{||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\lambda||^2}{n}.$$

If we let $\hat{\lambda}_n$ denote the regularization parameter selected by a given selection procedure, then the procedure is defined to be *asymptotically loss efficient* if

$$\frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_n})}{\inf_{\lambda \in [0, \lambda_{max}]} L(\hat{\boldsymbol{\beta}}_\lambda)} \to_p 1$$

and $\hat{\boldsymbol{\beta}}_{\hat{\lambda}_n}$ is said to be an *asymptotically loss efficient estimator.*

For the efficiency proofs we further require the following notation. In classical regression the risk function is defined as

$$R(\hat{\boldsymbol{\beta}}_\alpha) = \mathrm{E}_0\left(\frac{||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\alpha||^2}{n}\right) = \Delta_\alpha + \frac{\sigma^2 d_\alpha}{n},$$

where E_0 denotes expectation under the true model and $\Delta_\alpha = ||\boldsymbol{\mu} - \mathbf{H}_\alpha\boldsymbol{\mu}||^2/n$. Letting d_{α_λ} denote the number of predictors with non-zero coefficients in the penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$, we further define the function

$$\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) = \Delta_{\alpha_\lambda} + \frac{\sigma^2 d_{\alpha_\lambda}}{n},$$

which is a random variable.

2.1 Model Selection Procedures

K-fold CV is commonly used to select tuning parameters in both the statistical and machine learning literature. It operates by first randomly dividing the data set into k roughly equally sized subsets, then for each subset, the prediction error is computed based on the model fit using the data excluding that subset. The tuning parameter that minimizes the average square error computed across the subsets is then selected. In classical regression it has been shown that k-fold CV should have the same asymptotic properties as GIC_{κ_n} with

$$\kappa_n = \frac{2k-1}{k-1}$$

(Shao, 1997). Applying this result, 10-fold CV should have the same asymptotic performance as GIC_{κ_n} with $\kappa_n = 2.\overline{11}$, suggesting that 10-fold CV should be efficient. Under the assumption of an orthonormal design matrix Leng et al. (2006) showed that if the Lasso-estimated model minimizes the prediction error then it will fail to select the true model with non-zero probability. The authors noted that this suggests that k-fold CV is inconsistent, but to our knowledge, the asymptotic properties of k-fold CV have not been fully established in the context of penalized regression. While a rigorous extension of the classical theory for k-fold CV to penalized regression is beyond the scope of this paper, the simulation results suggest that the k-fold CV is efficient in the current context.

In addition to 10-fold CV, we study the performance of several information criteria. Specifically, we consider

$$AIC_\lambda = \log(\hat{\sigma}^2_\lambda) + 2\frac{df_\lambda}{n},$$

$$AIC_{c_\lambda} = \log(\hat{\sigma}^2_\lambda) + 2\frac{df_\lambda + 1}{n - df_\lambda - 2},$$

$$BIC_\lambda = \log(\hat{\sigma}^2_\lambda) + \log(n)\frac{df_\lambda}{n},$$

$$GCV_\lambda = \frac{\hat{\sigma}^2_\lambda}{(1 - df_\lambda/n)^2},$$

and

$$C_{p_\lambda} = \hat{\sigma}^2_\lambda + 2\frac{df_\lambda \tilde{\sigma}^2}{n}.$$

In the above we define

$$\hat{\sigma}^2_\lambda = \frac{||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda||^2}{n}$$

and

$$\tilde{\sigma}^2 = \frac{||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\bar{\alpha}}||^2}{n - d_n - 1}.$$

With the exception of 10-fold CV, all of the above model selection procedures require a definition of the effective degrees of freedom for the penalized regression method. In what follows, we use a heuristic definition and define the effective degrees of freedom to be the number of non-zero coefficients in $\hat{\boldsymbol{\beta}}_\lambda$ and denote this by d_{α_λ}. Zou et al. (2007) proved that the number of non-zero coefficients is an unbiased estimator of the degrees of freedom for the Lasso. For SCAD, Fan and Li (2001) proposed setting the degrees of freedom equal to the trace of the approximate linear projection matrix. Based on Proposition 1 from ZLT, our efficiency proofs would still hold if this alternate definition is used.

2.2 Efficiency Results

We show here that assuming that the true model is not in the set of candidate models, C_{p_λ}, AIC_λ, GCV_λ, and AIC_{c_λ} are efficient selectors of the regularization parameter. The dimension of the full model, d_n, is allowed to tend to infinity with n but it is assumed that $d_n/n \to 0$. The efficiency proofs operate under the same assumptions as those of ZLT, which are presented here for completeness:

(A1) $(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}$ exists and its largest eigenvalue is bounded by a constant number C.

(A2) $E\varepsilon_1^{4q} < \infty$, for some positive integer q.

(A3) The risks of the least squares estimators $\hat{\boldsymbol{\beta}}_\alpha$ satisfy

$$\sum_{\alpha \in \mathcal{A}_n} (nR(\hat{\boldsymbol{\beta}}_\alpha))^{-q} \to 0.$$

(A4)

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{||\mathbf{b}||^2}{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \to_p 0,$$

where $\mathbf{b}$ is a $d_n \times 1$ vector where $b_i = p'_\lambda(|\hat{\beta}_{\lambda i}|)sgn(\hat{\beta}_{\lambda_i})$ for all i such that $|\hat{\beta}_{\lambda i}| > 0$ and is equal to 0 otherwise.

The first three assumptions are common in the literature on model selection. Assumption (A1) requires the matrix of predictors to have full column rank and (A2) implies that efficiency can still apply even when penalized least squares is used but the true distribution of the error terms is not Gaussian. Assumption (A3) puts a restriction on how close the candidate models can be to the true model and precludes any scenario where the true model is included in the set of candidate models. The last assumption, (A4), is the only assumption that involves the penalty function and ZLT provided the following three sufficient conditions for the assumption to be satisfied.

(S1) $\sqrt{n}\lambda_{\max} < M_1$ for all n for some constant $M_1 > 0$.

(S2) For any θ, $p'(\theta) \leq M_2 \lambda$ for some constant $M_2 > 0$.

(S3) $n||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}\boldsymbol{\mu}||^2/d_n \to \infty$ as $n \to \infty$.

As pointed out by an anonymous referee, assumption (A3) restricts the size of the set of candidate models. The classical literature on model selection primarily worked with nested subsets and did not require the consideration of all subsets (e.g., Shibata (1981), Shao (1997), and Li (1987)); however, since the subsets selected by methods such as the Lasso or SCAD are data dependent, the set of candidate models is random and we cannot rule out any particular candidate model a priori. Therefore we need $\mathcal{A}_n$ to include all 2^{d_n} subsets in order to use the theory from classical model selection. Alternatively, if the data analyst can assume that the error terms are normally distributed then assumption (A3) can be replaced by a weaker assumption from Shibata (1981).

(A3*) For any $0 < \delta < 1$, $\sum_{\alpha \in \mathcal{A}_n} \delta^{nR(\hat{\beta}_\alpha)} \to 0$,

The following lemma details the restrictions on the behavior of d_n.

Lemma 2.1. *Assume that for all n sufficiently large*

$$||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}\boldsymbol{\mu}||^2 \geq k_1 n d_n^{k_2} \tag{2.1}$$

for some positive constant k_1 and some constant $k_2 \leq 0$. Then (A3) will hold if

$$\lim_{n\to\infty} \frac{d_n}{\log_2(n)} < q, \tag{2.2}$$

and (A3*) will hold if

$$\lim_{n\to\infty} n d_n^{k_2-1} = \infty. \tag{2.3}$$

The proof is presented in the appendix. This lemma shows that under (A3) d_n can at most grow logarithmically with n; however, polynomial growth rates are allowed under assumption (A3*) so long as $d_n = n^c$ for $c < \frac{1}{1-k_2}$. Specific values of k_2 are worked out for the simulation examples considered in Section 4.1.

The asymptotic efficiency of C_{p_λ} is given by the following result.

Theorem 2.1. *Assuming (A1)-(A4) hold and that $d_n/n \to 0$ as $n \to \infty$, the regularization parameter, $\hat{\lambda}_n$, selected by minimizing C_{p_λ} yields an asymptotically loss efficient estimator, $\hat{\beta}_n(\hat{\lambda}_n)$.*

To further establish the efficiency of AIC_λ, GCV_λ and AIC_{c_λ} we require the following two theorems. The first proves the efficiency of GIC_λ with the true error variance replaced by the estimated error variance based on the candidate model.

Theorem 2.2. *Assuming (A1)-(A4) hold and that $d_n/n \to 0$ as $n \to \infty$, the regularization parameter, $\hat{\lambda}_n$, selected by minimizing*

$$\Gamma_n(\lambda) = \hat{\sigma}^2_\lambda \left(1 + \frac{2d_{\alpha_\lambda}}{n}\right)$$

yields an asymptotically loss efficient estimator, $\hat{\beta}_{\hat{\lambda}_n}$. The same result holds under normality of the error terms with (A3) replacing (A3).*

Next, we prove that any procedure that is asymptotically equivalent to $\Gamma_n(\lambda)$ is also efficient.

Theorem 2.3. *Assuming (A1)-(A4) hold and that $d_n/n \to 0$ as $n \to \infty$, any information criterion that can be written in the form*

$$\tilde{\Gamma}_\lambda = \hat{\sigma}^2_\lambda \left(1 + \frac{2d_{\alpha_\lambda}}{n} + \delta_\lambda\right),$$

where

$$\sup_{\lambda \in [0, \lambda_{max}]} |\delta_\lambda| \to_p 0 \tag{C1}$$

and

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{|\delta_\lambda|}{L(\hat{\beta}_\lambda)} \to_p 0, \tag{C2}$$

is an asymptotically loss efficient procedure for selecting λ. The same result holds under normality of the error terms with (A3) replacing (A3).*

Condition (C2) in Theorem 2.3 is a stronger assumption than in the analogous result established by Theorem 4.2 in Shibata (1980) for selecting the optimal order of a linear process, but Theorem 2.3 is sufficient to show that AIC_λ, GCV_λ, and AIC_{c_λ} are asymptotically loss efficient model selection procedures for the regularization parameter. All three methods can be shown to satisfy (C1) and (C2) using Taylor series expansions. The details are provided in the supplementary material.

Remark 1. The efficiency proofs in this section make use of the results from Li (1987), which operate under assumptions (A1)-(A3). Similar results exist in Shibata (1981) if the error terms are normally distributed and (A3*) is substituted for (A3). The efficiency of AIC_λ, AIC_{c_λ}, and GCV_λ can be shown in a similar manner in this setting.

3 GLMs with No Dispersion Parameter

We now generalize our efficiency results to a broader class of models by studying the asymptotic performance of AIC_λ as a selector of λ when the likelihood function is misspecified as a generalized linear model (GLM) and prove that it is asymptotically loss efficient. We assume that the data $y_1, \ldots, y_n$ are independent with common unknown probability density function $g(y)$ and that $\mathrm{E}(y_i) = \mu_i$ and $\mathrm{Var}(y_i) = \sigma_i^2$. To approximate this distribution, we consider a family of GLMs where the density of each candidate model is given by

$$f_\alpha(y_i; \boldsymbol{\beta}_\alpha) = \exp\left(y_i \boldsymbol{\theta}_{\alpha i} - b(\boldsymbol{\theta}_{\alpha i}) + c(y_i)\right),$$

where $\boldsymbol{\theta}_\alpha = \mathbf{X}_\alpha \boldsymbol{\beta}_\alpha$, for $\alpha \in \mathcal{A}_n$. Here we have assumed that there is no dispersion parameter, and we further assume that $b(\theta)$ is three times differentiable and that $b''(\theta) > 0$ for all θ. All of these assumptions would hold for probit or logistic regression and the Poisson log-linear model.

A reasonable objective in this setting is to minimize two times the average Kullback-Leibler (KL) loss function, which is defined as

$$L_{KL}(\boldsymbol{\beta}_\alpha) = \frac{2}{n}\sum_{i=1}^{n} \mathrm{E}_0\left(\log g(y_i)\right) - \mathrm{E}_0\left(\log f_\alpha(y_i; \boldsymbol{\beta}_\alpha)\right) = \frac{2}{n}\sum_{i=1}^{n}\left[\mu_i(\boldsymbol{\theta}_{0i} - \boldsymbol{\theta}_{\alpha i}) + (b(\boldsymbol{\theta}_{\alpha i}) - b(\boldsymbol{\theta}_{0i}))\right].$$

For a given sample size n, we define $\boldsymbol{\theta}_\alpha^* = \mathbf{X}_\alpha \boldsymbol{\beta}_\alpha^*$ as the minimizer of the KL loss. By Theorem

1 in Lv and Liu (2010) we have that $\boldsymbol{\theta}^*_\alpha$ is the unique solution to the equation

$$\mathbf{X}'_\alpha(\boldsymbol{\mu} - b'(\boldsymbol{\theta})) = 0. \tag{3.1}$$

If $g(y) = f_\alpha(y; \boldsymbol{\beta}_0)$ for some true parameter $\boldsymbol{\beta}_0$ for any α, then $\boldsymbol{\beta}^*_\alpha = \boldsymbol{\beta}_0$. However, if we assume that we are in the non-true model world, then $g(y)$ is not completely specified by any of the candidate models and we refer to $\boldsymbol{\beta}^*_\alpha$ as the “pseudo-true parameter” based on the candidate model α.

Similarly to the Gaussian model, for a given λ, we take $\hat{\boldsymbol{\theta}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$ and denote the maximum-likelihood estimator based on the model α_λ by $\hat{\boldsymbol{\theta}}_{\alpha_\lambda} = \mathbf{X}_{\alpha_\lambda}\hat{\boldsymbol{\beta}}_{\alpha_\lambda}$. If we let $\hat{\lambda}_n$ denote the regularization parameter selected by a given selection procedure, then the procedure is defined to be *asymptotically loss efficient* if

$$\frac{L_{KL}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_n})}{\inf_{\lambda\in[0,\lambda_{max}]} L_{KL}(\hat{\boldsymbol{\beta}}_n(\lambda))} \to_p 1$$

and $\hat{\boldsymbol{\beta}}_n(\hat{\lambda}_n)$ is said to be an *asymptotically loss efficient estimator.*

ZLT studied the asymptotic performance of AIC_λ in a similar setting. To establish asymptotic loss efficiency, ZLT restricted the set of candidate models to the set

$$\mathcal{D} = \{\alpha : \sup_{\alpha\in\mathcal{D}} |\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_0| \to 0 \text{ in probability, as } n \to \infty\},$$

where $\boldsymbol{\theta}_0 = \mathbf{X}\boldsymbol{\beta}_0$. For this restricted set of models, the maximum-likelihood estimator converges uniformly to the true parameter. If this set is known in practice, then the model selection process reduces to selecting the most parsimonious model in this set. This class of models would rarely be known in practice, so this motivates us to weaken this assumption and to prove the efficiency of AIC_λ for a general set of candidate models.

Under the regularity conditions (R1)-(R2) given in the supplementary material, White (1982) proved that $\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha \to 0$, almost surely, and established the asymptotic normality of $\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha$ under (R1)-(R4). With the additional condition (R5), Nishii (1988) applied a Taylor expansion to show that

$$\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha = \mathbf{A}_n^{-1}\left\{\frac{1}{n}\frac{\partial l(\boldsymbol{\beta}^*_\alpha)}{\partial \boldsymbol{\beta}} + \mathbf{r}\right\} \tag{3.2}$$

for n sufficiently large, where $\mathbf{A}_n = -\frac{1}{n}\partial^2 l(\boldsymbol{\beta}^*_\alpha)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ and $r_j = O_p(||\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha||^2)$ for $j = 1, \ldots, d_\alpha$.

We define the risk function of the maximum-likelihood estimator to be $R_{KL}(\hat{\boldsymbol{\beta}}_\alpha) = \mathrm{E}_0(L_{KL}(\hat{\boldsymbol{\beta}}_\alpha))$. From Theorem 4 of Lv and Liu (2010), under (R1)-(R6),

$$R_{KL}(\hat{\boldsymbol{\beta}}_\alpha) = L_{KL}(\boldsymbol{\beta}^*_\alpha) + \frac{tr\{(\mathbf{X}_\alpha^T\mathbf{W}_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha^T\mathbf{W}_0\mathbf{X}_\alpha\}}{n} + o(1)$$

where $\mathbf{W}_0 = diag\{\sigma_1^2, \ldots, \sigma_n^2\}$ and $\mathbf{W}_\alpha = diag\{b''(\theta_{\alpha 1}), \ldots, b''(\theta_{\alpha n})\}$. Similarly to the Gaussian model, we further define the random variable

$$\tilde{R}_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) = L_{KL}(\boldsymbol{\beta}^*_{\alpha_\lambda}) + \frac{tr\{(\mathbf{X}_{\alpha_\lambda}^T\mathbf{W}_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda})^{-1}\mathbf{X}_{\alpha_\lambda}^T\mathbf{W}_0\mathbf{X}_{\alpha_\lambda}\}}{n} + o(1).$$

With these results and the following assumptions, we can prove the efficiency of AIC_λ.

(A1′) $(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}$ exists and its minimum and maximum eigenvalues are bounded below and above by constant numbers C_1 and C_2, respectively.

(A2′) $E(y_i - \mu_i)^{4q} < \infty$, for $i = 1, \ldots, n$ and some positive integer q.

(A3′) The risks of the maximum-likelihood estimators $\hat{\boldsymbol{\beta}}_\alpha$ satisfy

$$\sum_{\alpha\in\mathcal{A}_n}(nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha))^{-q} \to 0.$$

(A4′) $\sup_\theta b''(\theta) < \infty$

(A5′) $\sqrt{n}\lambda_{\max} < M_1$ for all n for some constant $M_1 > 0$.

(A6′) For any θ, $p'(\theta) \leq M_2\lambda$ for some constant $M_2 > 0$.

(A7′) $nL_{KL}(\beta^*_{\tilde{\alpha}})/d_n \to \infty$ as $n \to \infty$.

The first three assumptions are analogous to the assumptions made in the Gaussian model, and assumption (A4′) is a mild regularity assumption. As shown by the following lemma, assumptions (A5′)-(A7′) are sufficient conditions for the penalized estimator to be close to the maximum-likelihood estimator. These assumptions are analogous to the sufficient conditions used in the Gaussian model. They are stated explicitly here since they are required in parts of the efficiency proof.

Lemma 3.1. *Under (A5′)-(A7′),*

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{||\mathbf{b}||^2}{\tilde{R}_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \to_p 0,$$

where $\mathbf{b}_i$ *is a* $d_n \times 1$ *vector where* $b_i = p'_\lambda(|\hat{\beta}_{\lambda i})|)sgn(\hat{\beta}_{\lambda i})$ *for all* i *such that* $|\hat{\beta}_{\lambda i}| > 0$ *and is equal to* 0 *otherwise.*

The proof is given in Appendix B. The next theorem establishes the efficiency of AIC_λ.

Theorem 3.1. *Assuming* $d_n/n \to 0$ *as* $n \to \infty$*, (A1′)-(A7′) and the regularity conditions (R1)-(R6), the regularization parameter,* $\hat{\lambda}_n$*, selected by minimizing* AIC_λ *yields an asymptotically loss efficient estimator,* $\hat{\beta}_n(\hat{\lambda}_n)$*.*

The proof is given in Appendix B.

4 Simulation Studies

In this section we study the finite sample performance of the model selection procedures when the true model is not included in the set of candidate models.

In all of the examples, the results are based on 1000 realizations of samples with $n = 100, 200$, and 400, and the selection procedures are evaluated based on their loss efficiency, loss, and the variability of the selected number of non-zero coefficients. For each realization, if we let $\hat{\lambda}_n$ denote the regularization parameter selected by a given selection procedure, then the loss efficiency is computed as

$$\frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_n})}{\min_{\lambda \in [0,\lambda_{max}]} L(\hat{\boldsymbol{\beta}}_\lambda)}.$$

where $L(\cdot)$ is the L_2 loss in the linear regression examples and is the KL loss in the GLM examples. For comparison, we also include results for the (infeasible) "Optimal" procedure, which selects the tuning parameter over the regularization path that produces the minimum loss for each realization and report the loss ("Min.Loss") achieved by this procedure.

4.1 Linear Regression

In this section we study the finite sample performance of the model selection procedures discussed in Section 2.2. The first set of simulations considers a trigonometric regression where the candidate models are in the neighborhood of the true model but never include the true model. This example is in line with the framework considered by Shibata (1980) and Hurvich and Tsai (1991). The second set of simulations look at an example where there is an omitted predictor. For example, the researcher may have access to some of the relevant predictors but may be missing others. This is the setting that was considered by ZLT.

4.1.1 Choice of Penalty Function

We consider two common choices for the penalty function. The first is the Smoothly Clipped Absolute Deviation (SCAD) penalty function proposed by Fan and Li (2001). This penalty function is defined by

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\}$$

for some $a > 2$ and $\beta > 0$. Fan and Li (2001) recommended setting the second tuning parameter in the SCAD penalty function, a, equal to 3.7 and this is commonly done in practice; however, doing so will not necessarily guarantee that the SCAD objective function is convex and can result in convergence to local, but non-global, minima. As a result, in addition to studying the performance of SCAD with $a = 3.7$ (SCAD, 3.7), we study the performance of SCAD where $a = \max(3.7, 1 + 1/c^*)$ (SCAD) where c^* is the minimum eigenvalue of $n^{-1}\mathbf{X}'\mathbf{X}$. The latter choice will force the objective function to be convex (Breheny and Huang, 2011).

The wide use of SCAD is mainly due to the fact that it satisfies the "oracle property." This means that, assuming that the true model is in the set of candidate models and subject to certain regularity assumptions, there exists a sequence $\{\lambda_n\}$ such that if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ then with probability tending to one the SCAD-estimated regression based on the full model will correctly zero out any zero coefficients and have the same asymptotic distribution as the least squares regression based on the correct model. This result was proven originally for d_n fixed by Fan and Li (2001) and was extended to the case where $d_n < n$ but $d_n \to \infty$ by Fan and Peng (2004). These results are for an unknown deterministic sequence that needs to be estimated in practice.

The second penalty function that we study is the Lasso proposed by Tibshirani (1996). The Lasso penalty is the L_1-norm of the coefficients. Necessary and sufficient conditions have been established for the Lasso to perform consistent model selection (Zhao and Yu, 2006), but in general the Lasso produces biased estimates and does not satisfy the oracle property (Zou, 2006). However, in the non-true model world, the oracle property has no meaning, since there is no true model. Further, even in the true model world, the oracle property is an asymptotic property.

It is important to note that although ZLT only studied non-concave penalty functions, if the non-zero estimated coefficients, $\hat{\boldsymbol{\beta}}_{\lambda 1}$, satisfy a relationship of the form

$$\hat{\boldsymbol{\beta}}_{\lambda 1} = (\mathbf{X}'_{\alpha_\lambda}\mathbf{X}'_{\alpha_\lambda})^{-1}\mathbf{X}_{\alpha_\lambda}\mathbf{y} + \left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\mathbf{X}'_{\alpha_\lambda}\right)^{-1}\mathbf{b}_1$$

with probability tending to 1 and (A4) is satisfied, then the efficiency proofs will hold for any penalty function. In the above, $\mathbf{b}_1$ are the elements of $\mathbf{b}$ that correspond to $\hat{\boldsymbol{\beta}}_{\lambda 1}$. In particular, based on Lemma 2 of Zou et al. (2007), the Lasso satisfies this relationship and the same sufficient conditions provided by ZLT for (A4) can be used. Therefore, the efficiency proofs will hold for the Lasso, so it is interesting to compare the performance of the two penalty functions.

The Lasso regressions are fit using the R `lars` package (Hastie and Efron, 2011) and the SCAD regressions are fit using the R `ncvreg` package (Breheny and Huang, 2011). The `lars` package computes the entire regularization path for the Lasso and for SCAD the models are fit over a grid of 200 λ values from λ_{min} to λ_{max}, where the first 100 values of λ are fit on a log-scale and the last 100 values of λ are equally spaced. Breheny and Huang (2011) considered a grid of 100 λ values in their simulation studies. We have chosen a grid that is twice as fine in order to remain closer to the theoretical assumption that all possible values of λ are considered. In all simulations, λ_{max} is specified so that all of the estimated coefficients are zero and λ_{min} is chosen to effectively produce the least squares estimate on the full model.

4.1.2 Exponential model

Here we consider a trigonometric example based on an example studied in Hurvich and Tsai (1991). The true model is the model described as

$$y_i = e^{4i/n} + \varepsilon_i$$

for $i = 1, \ldots, n$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The estimated models are SCAD and Lasso penalized regressions where the matrix of predictors, $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2)$, is a $n \times d_n$ matrix with components defined by

$$x_{ij}^1 = \sin\left(\frac{2\pi j}{n} i\right)$$

and,

$$x_{ij}^2 = \cos\left(\frac{2\pi j}{n} i\right)$$

for $j = 1, \ldots, d_n/2$ and $i = 1, \ldots, n$. The maximum number of predictors is allowed to vary by letting the dimension $d_n = 2\lfloor n^c/2 \rfloor$. It is shown in the appendix that for this example $||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}||^2 \geq k_1 n d_n^{-2}$ for some positive constant k_1. Therefore, by Lemma 2.1, assumption (A3*) will hold so long as $c < 1/3$. In the simulations we take $c = .3$, and for comparison we also consider $c = .5$, .8 and .98. Note that examining d_n close to n allows for the study of high-dimensional data problems, and is in the spirit of simulations performed in Tibshirani (1996) and Zou and Hastie (2005). Since the predictor variables are orthogonal in this example, setting $a = 3.7$ for SCAD satisfies the convexity constraint for all values of c.

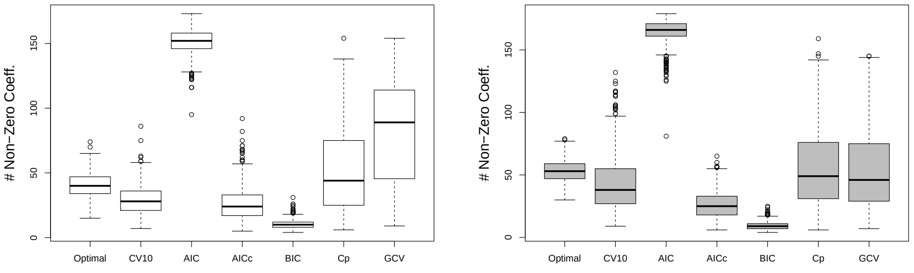
As in Hurvich and Tsai (1991), we examine both $\sigma^2 = 50$ and $\sigma^2 = 100$, but the patterns for the two error variances are similar so only the results for $\sigma^2 = 100$ are reported. The median L_2 loss efficiency is presented in Table 4.1.2 for both SCAD and Lasso. For all values of c, the median loss efficiency of AIC_{c_λ} and C_{p_λ} tend to one as the sample size increases, while the median loss efficiency of BIC_λ does not show signs of convergence. These patterns are consistent with the theoretical efficiency results. When the number of predictor variables is small relative to the sample size, the loss efficiency of AIC_λ also tends to one; however, as the number of candidate predictors is increased, the performance of AIC_λ deteriorates. Figure 1 displays boxplots of the selected number of non-zero coefficients when $n = 200$, $\sigma^2 = 100$, and $c = .98$. From this plot we see that AIC_λ often selects a model that is close to the full model when c is large. As the sample size is increased the full model becomes less desirable and AIC_λ suffers as a result. For SCAD, GCV_λ appears to suffer from a similar problem, but to a lesser extent than AIC_λ. The difference in performance for varying values of c suggests that the good asymptotic performance of AIC_λ and GCV_λ is strongly dependent on the fact that $d_n/n \to 0$ and these selectors may not perform well in finite samples when this ratio is close to 1.

Overall, the sensitivity to the value of c clearly hurts the performance of AIC_λ and can also negatively impact the performance of C_{p_λ} and GCV_λ. The impact on the latter two is more noticeable when looking at SCAD, but in both cases the extreme variability in the size of the selected model is undesirable. As a result, we recommend the use of AIC_c or 10-fold CV, which are less sensitive to the closeness of d_n to n.

Table 1: Median L2 Loss Efficiency over 1000 simulations for the exponential model with $\sigma^2 = 100$.

		Median Loss Efficiency							
		SCAD				Lasso			
Info. Crit.	n	c=.3	c=.5	c=.8	c=.98	c=.3	c=.5	c=.8	c=.98
10-fold CV	100	1.00	1.05	1.07	1.08	1.00	1.01	1.05	1.12
	200	1.00	1.03	1.06	1.05	1.00	1.01	1.03	1.07
	400	1.00	1.03	1.03	1.04	1.00	1.01	1.02	1.04
AIC_λ	100	1.00	1.04	1.18	2.43	1.00	1.01	1.07	2.13
	200	1.01	1.02	1.20	3.08	1.00	1.01	1.06	2.57
	400	1.00	1.02	1.23	4.05	1.00	1.01	1.05	3.29
AIC_{c_λ}	100	1.00	1.04	1.09	1.13	1.00	1.02	1.10	1.21
	200	1.01	1.03	1.07	1.08	1.00	1.01	1.06	1.11
	400	1.00	1.02	1.05	1.06	1.00	1.01	1.04	1.08
BIC_λ	100	1.00	1.07	1.32	1.64	1.00	1.05	1.60	1.64
	200	1.02	1.06	1.47	1.51	1.00	1.06	1.74	1.62
	400	1.01	1.07	1.60	1.51	1.00	1.08	1.80	1.60
C_{p_λ}	100	1.00	1.04	1.10	1.22	1.00	1.01	1.05	1.15
	200	1.01	1.02	1.09	1.15	1.00	1.01	1.03	1.09
	400	1.00	1.02	1.08	1.09	1.00	1.01	1.03	1.05
GCV_λ	100	1.00	1.04	1.10	1.69	1.00	1.01	1.06	1.16
	200	1.01	1.02	1.10	1.73	1.00	1.01	1.04	1.09
	400	1.00	1.02	1.08	1.82	1.00	1.01	1.03	1.05

Figure 1: Comparison of model selection procedures based on the number of non-zero coefficients (includes intercept) in the selected model over 1000 simulations for the exponential model with $n = 200$, $\sigma^2 = 100$, and $c = 0.98$.



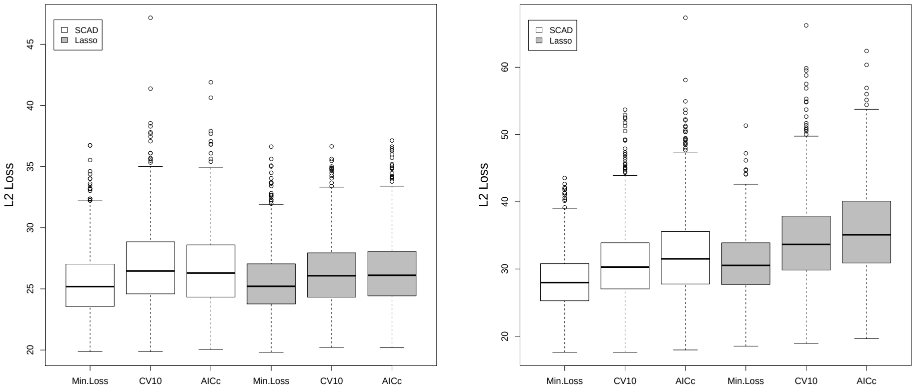
(a) SCAD

(b) Lasso

Figure 2 presents boxplots of the L_2 loss for the 1000 realizations when $n = 200$ when $c = .5$ and $c = .98$. From this we can compare the optimal performance of SCAD and the

Lasso. Based on minimum loss, the predictive accuracies of the two methods are similar. This reinforces that the existence of an oracle property is not relevant in the non-true model world, and an estimator that does not possess the oracle property can still be effective from a predictive point of view.

Figure 2: Comparison of model selection procedures based on L2 Loss over 1000 simulations for the exponential model with $n = 200$ and $\sigma^2 = 100$.



(a) c=.5

(b) c=.98

4.1.3 Omitted Predictor

Here we study an omitted predictor example similar to example 2 in ZLT. The true model is defined as

$$y_i = 3x_{i,1} + 1.5x_{i,2} + 2x_{i,10} + x_{i,13} + \varepsilon_i$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for $\sigma^2 = 16$ and $\sigma^2 = 25$. We let $\mathbf{X}$ be a $2n \times (d_n + 1)$ matrix of predictors where the $\mathbf{x}_i's$ are simulated from a multivariate normal distribution with mean 0 and variance-covariance matrix Σ where $\Sigma_{i,j} = \rho^{|i-j|}$ for $\rho = 0$ and 0.5. In the simulations $\boldsymbol{X}$ is simulated once and is used for every simulation run in order to resemble a fixed $\boldsymbol{X}$ setting. The estimated models are SCAD and Lasso penalized regressions based on the first n observations of $\mathbf{X}$ except with the 13^{th} column removed so that the true model is never included in the set of candidate models. In order to compare predictive performance, we treat the remaining observations of $\boldsymbol{X}$ as a hold-out sample and use it to compute the loss for each estimated model.

In both examples the number of superfluous variables included in the candidate models is allowed to vary by letting the dimension $d_n = 2\lfloor n^c/2 \rfloor$. Under deterministic $\boldsymbol{X}$, it is shown in the supplementary material that $||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}||^2 \geq k_1 n$ for some positive constant k_1 if the excluded predictor is orthogonal to the included predictors. By Lemma 2.1, assumption (A3$'$) will then hold if $d_n/n \to 0$. This suggests that when the excluded predictor is uncorrelated or only moderately correlated with the included predictors it is reasonable to compare $c = 0.5, 0.8$ and 0.98.

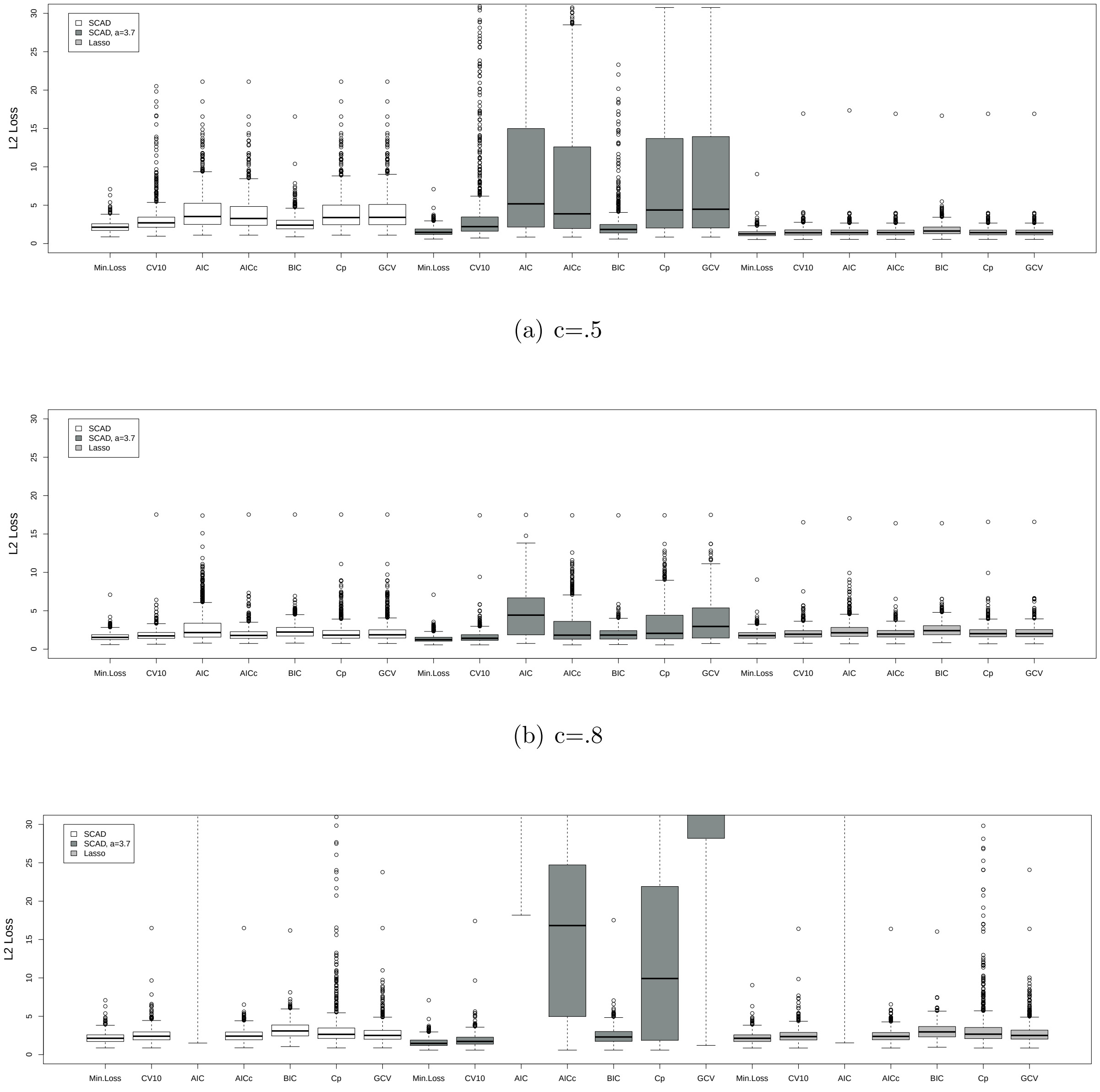
In this example setting $a = 3.7$ will not satisfy the convexity constraint for all values of c. Therefore, we further compare the case where $a = 3.7$ (SCAD, $a = 3.7$) to the case where $a = \max(3.7, 1 + 1/c^*)$ (SCAD).

The patterns for the two error variances and two values of ρ are similar so only the results for $\sigma^2 = 16$ and $\rho = 0.5$ are reported. We first consider Figure 3, which presents boxplots comparing the three estimators based on loss when $n = 200$. From these plots it is immediately clear that all of the information criteria perform better when a is allowed to be data-dependent, while 10-fold CV performs well regardless of the choice of a. One possible explanation for this is that all of the information criteria under consideration were derived for use in classical least squares regression so they should perform well assuming that the estimated models are close to the corresponding OLS models. When the second tuning parameter of SCAD is fixed at 3.7, the objective function is not necessarily convex so the SCAD-estimated models may be very far from the OLS models. On the other hand, 10-fold CV is a general model selection procedure that should work in a variety of settings. In general, we recommend using a data-dependent choice of a since it requires little additional cost and can greatly improve the performance of all of the information criteria.

Focusing only on the data-dependent choice of a, we see that the performance of the model selection procedures is similar for both SCAD and Lasso when $c = .8$ and when $c = .98$, but that the performance of SCAD is noticeably worse when $c = .5$. A possible explanation for this is that when c is small, the performance of the SCAD estimators is more sensitive to the choice of the second tuning parameter. Although taking $a = \max(3.7, 1 + 1/c^*)$ guarantees that the penalized loss function is convex, it may not be the optimal choice for this parameter and more investigation into the choice of this parameter is needed. Of course, this implies an advantage of Lasso over SCAD, since it does not require the choice of this second parameter.

Comparing the model selection procedures, we again see that AIC_λ, GCV_λ, and C_{p_λ} are sensitive to the number of predictor variables while AIC_{c_λ} and 10-fold CV maintain good performance. The boxplots of the selected number of non-zero coefficients are omitted since the patterns are similar to those seen in the exponential model. In Figure 3 it is clear that this sensitivity to the value of c impacts the performance of the model selection procedures, and as a result 10-fold CV and AIC_{c_λ} outperform the other procedures. 10-fold CV outperforms AIC_{c_λ} in some scenarios, but, in general, the performance of the two methods appears to be comparable.

Figure 3: Comparison of model selection procedures based on L2 Loss on new design points over 1000 simulations for the model with an omitted predictor with $n = 200$ and $\rho = 0.5$. In order to make it easier to compare the procedures, the limits of the vertical axis are specified so that all the boxes and whiskers appear but some of the outliers are not shown.



(a) c=.5

(b) c=.8

(c) c=.98

In order to study the asymptotic behavior of the selection procedures, Table 2 presents the median loss efficiencies. With the exception of SCAD with $c = 0.5$, the loss efficiencies of AIC_{c_λ}, C_{p_λ}, and GCV_λ tend to one, while the loss efficiency of BIC_λ does not show signs of convergence. Also, the results again show that AIC_λ performs poorly when the number of predictor variables is large relative to the sample size. For SCAD with $c = 0.5$, the loss efficiency of the efficient methods do not show signs of converging to one, which further suggests that the second tuning parameter may not be optimally selected. Overall, the results corroborate the theoretical findings, but reinforce that the finite sample performance of asymptotically equivalent methods may vary greatly.

Table 2: Median L2 Loss Efficiency on new design points over 1000 simulations for the model with an omitted predictor with $\rho = 0.5$.

		Median Loss Efficiency								
		SCAD			SCAD, a=3.7			Lasso		
Info. Crit.	n	c=.5	c=.8	c=.98	c=.5	c=.8	c=.98	c=.5	c=.8	c=.98
10-fold CV	100	1.32	1.10	1.09	1.72	1.23	1.20	1.08	1.09	1.08
	200	1.19	1.07	1.07	1.35	1.08	1.10	1.05	1.06	1.05
	400	1.14	1.05	1.05	1.26	1.02	1.04	1.04	1.04	1.04
AIC_λ	100	1.49	1.44	41.44	2.78	2.57	37.24	1.08	1.19	37.64
	200	1.57	1.24	51.80	3.03	3.30	59.94	1.06	1.12	49.77
	400	1.84	1.11	67.73	4.13	3.16	76.07	1.04	1.07	64.94
AIC_{c_λ}	100	1.36	1.13	1.10	2.19	1.45	4.27	1.07	1.09	1.08
	200	1.41	1.08	1.07	2.45	1.27	10.10	1.06	1.07	1.06
	400	1.68	1.06	1.05	3.31	1.10	17.28	1.04	1.04	1.05
BIC_λ	100	1.12	1.26	1.40	1.26	1.41	1.62	1.11	1.24	1.31
	200	1.07	1.39	1.40	1.13	1.31	1.38	1.21	1.34	1.33
	400	1.05	1.31	1.32	1.07	1.16	1.25	1.21	1.28	1.27
C_{p_λ}	100	1.42	1.17	1.22	2.40	1.65	3.04	1.08	1.12	1.24
	200	1.46	1.10	1.16	2.69	1.42	5.27	1.06	1.08	1.14
	400	1.75	1.07	1.08	3.75	1.14	10.93	1.04	1.05	1.08
GCV_λ	100	1.43	1.20	1.13	2.49	2.01	14.24	1.07	1.11	1.12
	200	1.48	1.11	1.10	2.75	2.10	25.71	1.06	1.08	1.09
	400	1.78	1.07	1.06	3.86	1.26	35.72	1.04	1.05	1.06

4.2 Poisson Regression

In this section we present simulation results for GLMs with no dispersion parameter. For GLMs, it is less clear how to handle the second tuning parameter for SCAD. Breheny and Huang (2011) recommended using an adaptive rescaling technique, but it is unclear how such a procedure will impact the performance of the model selection procedures and initial simulations for Bernoulli data resulted in convergence issues. As a result we only study the Lasso in this section. The `lars` package is only designed for linear regression, so we instead work with the R `glmpath` package (Park and Hastie, 2011), which fits the entire regularization path for the Lasso for GLMs.

We consider a trigonometric example based on an example studied in Hurvich and Tsai (1991). We take $\theta_t = e^{-5i/n}$ for $t = 0, \dots, n-1$ and simulate y_t from a Poisson distribution with $\mu_t = \exp(\theta_t)$. The estimated models are Lasso penalized Poisson regressions where the matrix of predictors, $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2)$, is a $n \times d_n$ matrix with components defined as in the exponential model. Similar to before, we vary the maximum number of predictors by letting the dimension $d_n = 2\lfloor n^c/2 \rfloor$ and we compare $c = .3$, $c = .5$ and $c = .8$. The case with $c = .98$ is omitted due to convergence problems with the package.

Although AIC_c was originally derived for linear regression, its use is commonly recommended in a more general setting when the number of predictor variables is large relative to the sample size (Burnham and Anderson (2002), p. 66). We therefore compare the performance of AIC_λ to 10-fold CV, AIC_{c_λ} and BIC_λ where

$$AIC_{c_\lambda} = -\frac{2}{n} l(\hat{\boldsymbol{\beta}}_\lambda) + 2\frac{df_\lambda + 1}{n - df_\lambda - 2}$$

and

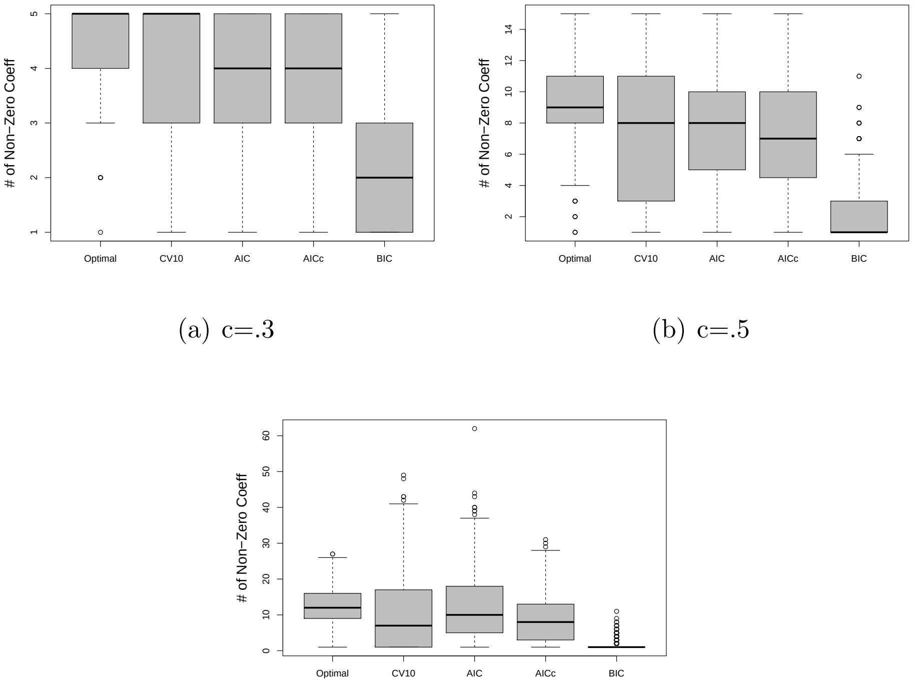
$$BIC_\lambda = -\frac{2}{n} l(\hat{\boldsymbol{\beta}}_\lambda) + \log(n)\frac{df_\lambda}{n}.$$

Table 4.2 reports the median KL loss efficiencies over the 1000 simulations. In all three cases, AIC_λ, AIC_{c_λ}, and 10-fold CV show signs of converging to one and have comparable performance, whereas BIC_λ performs noticeably worse and does not show signs of convergence. Figure 4 presents boxplots of the selected number of non-zero coefficients. This figure suggests that the poor performance of BIC_λ is due to its tendency to select models that are too sparse. In comparison, the other procedures select models with dimension closer to the optimal dimension. Overall, these results are consistent with the theoretical findings.

Table 3: Median KL Loss Efficiency over 1000 simulations for the poisson model.

		Median Loss Efficiency		
		Lasso		
Info. Crit.	n	c=.3	c=.5	c=.8
10-fold CV	100	1.08	1.30	1.17
	200	1.00	1.19	1.17
	400	1.00	1.08	1.09
AIC_λ	100	1.01	1.19	1.15
	200	1.01	1.13	1.10
	400	1.01	1.08	1.08
AIC_{c_λ}	100	1.02	1.18	1.10
	200	1.01	1.13	1.07
	400	1.01	1.08	1.06
BIC_λ	100	1.38	1.38	1.14
	200	1.48	1.63	1.27
	400	1.30	1.85	1.45

Figure 4: Comparison of model selection procedures based on the number of non-zero coefficients (includes intercept) in the selected model over 1000 simulations for the poisson model with $n = 200$



(a) c=.3

(b) c=.5

(c) c=.8

5 Analysis of a Real Data Set

We now consider the transaction data set from Sela and Simonoff (2012) in order to compare the candidate models chosen by the regularization parameter selectors when applied to a real world data set. The data contains transactions for third-party sellers on Amazon Web Services and the goal is to predict the prices at which software titles are sold based on the characteristics of the competing sellers. The target variable is the price premium that a seller can command (the difference between the price at which the good is sold and the average price of all of the competing goods in the marketplace). There are 24 potential predictors which include the seller's reputation (the total number of comments and the number of positive and negative comments received from buyers over different time periods), the length of time that the seller has been in the marketplace, the number of competitors, the quality of competing goods in the marketplace, the average reputation of the competitors, and the average prices of the competing goods. The data set contains 100 observations.

Table 5 reports the results for the information criteria as well as 10-fold CV based on two different runs (and hence two different random divisions of the data), which are referred to as 10-fold CV (1) and 10-fold CV (2). Only six predictor variables were ever selected so the remaining variables are omitted from the table. It is clear that the variables selected are heavily reliant on the selection procedure and the penalty function chosen. In particular, there is a noticeable difference between the variables selected by AIC_λ and AIC_{c_λ}, and in all three cases BIC_λ selected a model with no predictors, suggesting that it may be selecting an underfitted model. If we approach this problem from a predictive point of view, we know that there is little advantage to using SCAD over the Lasso, but that the choice of the second tuning parameter can greatly impact the performance of the former. Therefore, we recommend focusing on the Lasso. From the simulations we know that 10-fold CV maintains good performance in a variety of settings. However, it is 10 times more expensive to implement than using an information criterion, the asymptotic properties of 10-fold CV are not fully understood in this context, and the randomness involved in the procedure makes it difficult for data analysts to reproduce results. In the case of the Lasso, this last point is reinforced by the change in the selected variables between the two runs of 10-fold CV, as in the first run four nonzero coefficients were estimated, while in the second run none were. We recommend proceeding using AIC_{c_λ} as the selector of the tuning parameter for the Lasso as an alternative that avoids these issues.

Table 4: Selected variables for transaction data.

Selector	Ave. Comp. Price	Ave. Comp. Condition	Ave. Comp. Rating	Seller Condition	Negative Comments (30 days)	Negative Comments (Lifetime)
			SCAD			
10-fold CV (1)	X					
10-fold CV (2)	X					
AIC_λ	X	X	X	X	X	
AIC_{c_λ}	X					
BIC_λ						
C_{p_λ}	X					
GCV_λ	X	X	X	X	X	
			SCAD ($a = 3.7$)			
10-fold CV (1)	X					
10-fold CV (2)	X					
AIC_λ	X	X	X	X	X	X
AIC_{c_λ}	X	X	X	X	X	X
BIC_λ						
C_{p_λ}	X	X	X	X	X	X
GCV_λ	X	X	X	X	X	X
			LASSO			
10-fold CV (1)	X	X		X	X	
10-fold CV (2)						
AIC_λ	X	X	X	X	X	
AIC_{c_λ}	X					
BIC_λ						
C_{p_λ}	X					
GCV_λ	X					

6 Concluding Remarks

This paper studied the asymptotic and finite sample performance of classical model selection procedures in the context of penalized likelihood estimators without the assumption that the true model is included amongst the candidate models. We proved that AIC_λ, AIC_{c_λ}, C_{p_λ}, and GCV_λ are efficient selectors of the regularization parameter for regularized regression, and the numerical studies for regularized regression yielded several interesting observations. As anticipated, we found that BIC_λ is outperformed by the efficient model selection procedures and demonstrated that AIC_λ, BIC_λ, C_{p_λ}, and GCV_λ are all sensitive to the number of predictor variables that are included in the full model and that their performance can suffer as a result. In light of this issue we recommend that researchers use a method that is insensitive to the number of variables included in the model. From the simulations, 10-fold CV has the best overall performance. However, the discussion in Section 5 noted some of the disadvantages of this method including computational cost and variable results due to the inherent randomness of the procedure. As an alternative, data analysts can consider using AIC_{c_λ}, which was shown here to be an efficient selection procedure for the tuning parameter, and which the simulations suggest has comparable performance to that of 10-fold CV. Lastly, the simulations suggest that there is no clear advantage to using SCAD in a world where the "oracle property" does not apply. Combining this with the facts that the Lasso can be fitted using the efficient 'Lars' algorithm and does not involve a second tuning parameter that can greatly impact results, researchers may prefer to use the Lasso if they feel that they are in the non-true model world.

To further generalize our results, we also proved that AIC_λ is an efficient selector of the regularization parameter for regularized GLMs with no dispersion parameter and used numerical studies to compare its performance to that of AIC_{c_λ}, BIC_λ and 10-fold CV. Again, the performance of BIC_λ was noticeably worse than the other procedures, and the performances of AIC_λ, AIC_{c_λ} and 10-fold CV were comparable to each other, supporting our recommendation for the use of AIC_{c_λ}. Extending these results to GLMs with an unknown dispersion parameter is an interesting open problem. In this setting it is necessary to work with extended quasi-likelihood methods. Although model selection criteria such as AIC_c have been proposed in such settings as (Hurvich and Tsai, 1995), the extended quasi-likelihood is not a true likelihood so the results of White (1982) and Nishii (1988) do not apply. Investigations into the properties of model selection procedures in this setting is an area for future research.

As a final remark, this paper dealt with the case when $d_n/n \to 0$, and the theoretical results cannot be directly extended to the case when d_n/n converges to something other than zero. The latter setting has received a great deal of attention in recent literature (in particular $d_n \gg n$) and is an area for future investigation.

Appendix A

Proof of Lemma 2.1. By definition

$$nR(\hat{\boldsymbol{\beta}}_\alpha) \geq ||\boldsymbol{\mu} - \boldsymbol{H}_\alpha \boldsymbol{\mu}||^2 \geq ||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}} \boldsymbol{\mu}||^2.$$

Then by (2.1) and (2.2),

$$\sum_{\alpha=1}^{2^{d_n}} (nR(\hat{\beta}_\alpha))^{-q} \leq 2^{d_n} k_1^{-q} n^{-q} d_n^{-qk_2} = 2^{log_2(n)\left(\frac{d_n}{\log_2(n)} - \frac{qlog_2(k_1)}{log_2(n)} - \frac{qk_2 \log_2(d_n)}{log_2(n)}\right)} \to 0.$$

Next by (2.1) and (2.3),

$$\sum_{\alpha=1}^{2^{d_n}} \delta^{nR(\hat{\beta}_\alpha)} \leq 2^{d_n} \delta^{k_1 n^{k_2}} = 2^{d_n\left(1 + \frac{k_1 n d_n^{k_2} log_2(\delta)}{d_n}\right)} \to 0.$$

□

Before proving Theorems 1, 2, and 3, we establish the following two lemmas.

Lemma A.1. *Assume that (A1)-(A4) hold and that $d_n/n \to 0$ as $n \to \infty$. Then*

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{d_{\alpha_\lambda} |\hat{\sigma}_\lambda^2 - \sigma^2|}{nL(\hat{\beta}_\lambda)} \to_p 0.$$

Proof. The technique used to prove this result is similar to the proof of Theorem 2 in Shibata (1981). First consider

$$\begin{aligned}
|\hat{\sigma}_\lambda^2 - \sigma^2| &= \left| \frac{||\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda||^2}{n} - \sigma^2 \right| \\
&\leq \left| \frac{||\mathbf{y} - \hat{\boldsymbol{\mu}}_{\alpha_\lambda}||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2}{n} \\
&\leq L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) + 2 \left| \frac{\boldsymbol{\varepsilon}^T (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\alpha_\lambda})}{n} \right| + \left| \frac{||\boldsymbol{\varepsilon}||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2}{n}.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality, it follows that

$$|\hat{\sigma}^2_\lambda - \sigma^2| \leq L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) + 2||\boldsymbol{\varepsilon}|| \frac{||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\alpha_\lambda}||}{n} + \left| \frac{||\boldsymbol{\varepsilon}||^2}{n} - \sigma^2 \right| + \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2}{n}.$$

Then

$$\begin{aligned}
\frac{|\hat{\sigma}^2_\lambda - \sigma^2| d_{\alpha_\lambda}}{nL(\hat{\boldsymbol{\beta}}_\lambda)} \leq{}& \frac{d_{\alpha_\lambda}}{n} \left[\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \right] \\
&+ \frac{2}{\sigma} \left[\frac{d_{\alpha_\lambda}}{n} \frac{||\boldsymbol{\varepsilon}||^2}{n} \right]^{1/2} \left[\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \right]^{1/2} \left[\frac{\sigma^2 d_{\alpha_\lambda}}{n\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \right]^{1/2} \\
&+ \left[\frac{\sigma^2 d_{\alpha_\lambda}}{n\tilde{R}(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda))} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \right] \left| \frac{||\boldsymbol{\varepsilon}||^2}{n\sigma^2} - 1 \right| + \frac{d_{\alpha_\lambda}}{n} \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2}{nL(\hat{\boldsymbol{\beta}}_\lambda)}.
\end{aligned}$$

By definition,

$$\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) \geq \frac{\sigma^2 d_{\alpha_\lambda}}{n}.$$

Thus

$$\begin{aligned}
\frac{|\hat{\sigma}^2_\lambda - \sigma^2| d_{\alpha_\lambda}}{nL(\hat{\boldsymbol{\beta}}_\lambda)} \leq{}& \sup_{\lambda \in [0, \lambda_{max}]} \frac{d_{\alpha_\lambda}}{n} \left[\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \right] \\
&+ \sup_{\lambda \in [0, \lambda_{max}]} \frac{2}{\sigma} \left[\frac{d_{\alpha_\lambda}}{n} \frac{||\boldsymbol{\varepsilon}||^2}{n} \right]^{1/2} \left[\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \right]^{1/2} \left[\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \right]^{1/2} \\
&+ \sup_{\lambda \in [0, \lambda_{max}]} \left[\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \right] \left| \frac{||\boldsymbol{\varepsilon}||^2}{n\sigma^2} - 1 \right| + \sup_{\lambda \in [0, \lambda_{max}]} \frac{d_{\alpha_\lambda}}{n} \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2}{nL(\hat{\boldsymbol{\beta}}_\lambda)}.
\end{aligned} \tag{A.1}$$

Li (1987) established that

$$\sup_{\alpha \in \mathcal{A}_n} \left| \frac{L(\hat{\boldsymbol{\beta}}_\alpha)}{R(\hat{\boldsymbol{\beta}}_\alpha)} - 1 \right| \to_p 0$$

and it follows that

$$\sup_{\alpha \in \mathcal{A}_n} \left| \frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} - 1 \right| \to_p 0. \tag{A.2}$$

In addition, from the proof of Theorem 2 in ZLT we have that

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| \to_p 0, \tag{A.3}$$

and

$$\sup_{\lambda\in[0,\lambda_{max}]}\frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda||^2}{nL(\hat{\boldsymbol{\beta}}_\lambda)} \to_p 0. \tag{A.4}$$

Combining these results with the Law of Large Numbers and the assumption that $d_n/n \to 0$ as $n\to\infty$ the four terms on the right-hand side of equation (A.1) converge to 0 in probability. Hence,

$$\sup_{\lambda\in[0,\lambda_{max}]}\frac{2d_{\alpha_\lambda}|\hat{\sigma}^2_\lambda-\sigma^2|}{nL(\hat{\boldsymbol{\beta}}_\lambda)} \to_p 0$$

as desired. □

Lemma A.2. *Assume that (A1)-(A4) hold and that $d_n/n \to 0$ as $n\to\infty$. Then*

$$\sup_{\lambda\in[0,\lambda_{max}]}\frac{d_{\alpha_\lambda}|\tilde{\sigma}^2_n-\sigma^2|}{nL(\hat{\beta}_\lambda)} \to_p 0.$$

Proof. Start by noting that for all $\lambda\in[0,\lambda_{max}]$, $\Delta_{\alpha_\lambda}\geq\Delta_{\bar{\alpha}}$ (ZLT). Consider

$$\begin{aligned}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})d_{\alpha_\lambda}}{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})d_n} &\leq \frac{(\Delta_{\bar{\alpha}}+\frac{d_n\sigma^2}{n})d_{\alpha_\lambda}}{(\Delta_{\bar{\alpha}}+\frac{d_{\alpha_\lambda}\sigma^2}{n})d_n}\\ &\leq \frac{\Delta_{\bar{\alpha}}}{\Delta_{\bar{\alpha}}+\frac{d_{\alpha_\lambda}\sigma^2}{n}}+\frac{\frac{d_n\sigma^2}{n}d_{\alpha_\lambda}}{\frac{d_{\alpha_\lambda}\sigma^2}{n}d_n(\bar{\alpha})}\\ &\leq 2.\end{aligned}$$

From the proof of Lemma 1 we have that

$$|\tilde{\sigma}^2_n-\sigma^2| \leq \frac{n}{n-d_n-1}L(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})+2\frac{n}{n-d_n-1}||\boldsymbol{\varepsilon}_n||\frac{||\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{\bar{\alpha}}||}{n}+\left|\frac{||\boldsymbol{\varepsilon}||^2}{n-d_n-1}-\sigma^2\right|.$$

Thus

$$\frac{d_n|\tilde{\sigma}^2-\sigma^2|}{nL(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})} \leq \frac{n}{n-d_n-1}\frac{d_n}{n} + \frac{2}{\sigma}\frac{n}{n-d_n-1}\left[\frac{||\boldsymbol{\varepsilon}||^2}{n}\frac{d_n}{n}\right]^{1/2}\left[\frac{\sigma^2 d_n}{n\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}{L(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\right]^{1/2}$$
$$\left[\frac{d_n\sigma^2}{n\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}{L(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\right]\left|\frac{||\boldsymbol{\varepsilon}||^2}{(n-d_n-1)\sigma^2}-1\right|.$$

Under the assumption that $d_n/n \to 0$ as $n \to \infty$ it follows that

$$\frac{d_n|\tilde{\sigma}^2-\sigma^2|}{nL(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})} \to_p 0.$$

Combining these results with (A.2) and (A.3) it follows that

$$\begin{aligned}\frac{d_{\alpha_\lambda}|\tilde{\sigma}^2-\sigma^2|}{nL(\hat{\beta}_\lambda)} &\leq \sup_{[0,\lambda_{max}]}\frac{d_n|\tilde{\sigma}^2-\sigma^2|}{nL(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\frac{d_{\alpha_\lambda}\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}{d_n\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}\frac{L(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}{\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)}\\ &\leq 2\frac{d_n|\tilde{\sigma}^2-\sigma^2|}{nL(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\sup_{[0,\lambda_{max}]}\frac{L(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}{\tilde{R}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}})}\frac{\tilde{R}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} \to_p 0.\end{aligned}$$

□

Proof of Theorem 1. As in the proofs in ZLT, to prove that C_{p_λ} is asymptotically loss efficient, it is sufficient to show that

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{C_{p_\lambda}-||\boldsymbol{\varepsilon}||^2/n-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| \to_p 0. \qquad \text{(A.5)}$$

Decomposing C_{p_λ} it can be established that

$$\begin{aligned}C_{p_\lambda} &= \frac{||\mathbf{y}-\hat{\boldsymbol{\mu}}_\lambda||^2}{n} + \frac{2\tilde{\sigma}^2 d_{\alpha_\lambda}}{n}\\ &= \frac{||\boldsymbol{\varepsilon}||^2}{n} + L(\hat{\boldsymbol{\beta}}_\lambda) + (L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})-L(\hat{\boldsymbol{\beta}}_\lambda)) + \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda||^2}{n}\\ &+ \frac{2\boldsymbol{\varepsilon}^T[I-\mathbf{H}_{\alpha_\lambda}]\boldsymbol{\mu}}{n} + \frac{2(\sigma^2 d_{\alpha_\lambda}-\boldsymbol{\varepsilon}^T\mathbf{H}_{\alpha_\lambda}\boldsymbol{\varepsilon})}{n} + \frac{2(\tilde{\sigma}^2-\sigma^2)d_{\alpha_\lambda}}{n}.\end{aligned}$$

The proof of Theorem 2 in ZLT established that

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{2\boldsymbol{\varepsilon}^T(I-\mathbf{H}_{\alpha_\lambda})\boldsymbol{\mu}}{nL(\hat{\boldsymbol{\beta}}_\lambda)}\right|\rightarrow_p 0,$$

and,

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{2(\sigma^2 d_{\alpha_\lambda}-\boldsymbol{\varepsilon}_n^T\mathbf{H}_{\alpha_\lambda}\boldsymbol{\varepsilon})}{nL(\hat{\boldsymbol{\beta}}_\lambda)}\right|\rightarrow_p 0.$$

Combining these results with (A.2)-(A.4) and Lemma 2, (A.5) follows as desired. □

Proof of Theorem 2. The proof is the same as that of Theorem 1 except that the estimated variance is based on the candidate model rather than the full model and the result is established by using Lemma 1 in place of Lemma 2. □

Proof of Theorem 3. As in the efficiency proof for Γ_λ, it is sufficient to show that

$$\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\tilde{\Gamma}_\lambda-||\boldsymbol{\varepsilon}||^2/n-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right|\rightarrow_p 0 \tag{A.6}$$

to establish that $\tilde{\Gamma}_\lambda$ is an asymptotically efficient selection procedure for the regularization parameter, λ. By the definition of $\tilde{\Gamma}_\lambda$ we have that

$$\begin{aligned}\sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\tilde{\Gamma}_\lambda-||\boldsymbol{\varepsilon}||^2/n-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| &= \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\delta_\lambda\hat{\sigma}_\lambda^2+\Gamma_\lambda-||\boldsymbol{\varepsilon}||^2/n-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right|\\ &\leq \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\delta_\lambda(\hat{\sigma}_\lambda^2-\sigma^2)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right|+\sup_{\lambda\in[0,\lambda_{max}]}\frac{|\delta_\lambda|\sigma^2}{L(\hat{\boldsymbol{\beta}}_\lambda)}\\ &+ \sup_{\lambda\in[0,\lambda_{max}]}\left|\frac{\Gamma_\lambda-||\boldsymbol{\varepsilon}||^2/n-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right|.\end{aligned}$$

The last two terms converge to zero by (C1) and the efficiency proof for Γ_λ. From the proof of the previous lemma we further have that

$$\left|\frac{\delta_\lambda(\hat{\sigma}^2_\lambda - \sigma^2)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| \leq |\delta_\lambda| \frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)} + 2\frac{||\boldsymbol{\varepsilon}||}{\sqrt{n}} \left(\frac{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right)^{1/2} \left(\frac{|\delta_\lambda|}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right)^{1/2} (|\delta_\lambda|)^{1/2}$$
$$+ \frac{|\delta_\lambda|}{L(\hat{\boldsymbol{\beta}}_\lambda)} \left|\frac{||\boldsymbol{\varepsilon}||^2}{n} - \sigma^2\right| + |\delta_\lambda| \frac{||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2}{nL(\hat{\boldsymbol{\beta}}_\lambda)}.$$

By (C1), (C2), and similar arguments as those used in the efficiency proof for Γ_λ we have that the right hand side converges to 0 in probability. Therefore, it follows that

$$\sup_{\lambda\in[0,\lambda_{max}]} \left|\frac{\delta_\lambda(\hat{\sigma}^2_\lambda - \sigma^2)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| \to_p 0$$

and so equation (A.6) holds as desired. □

Appendix B

Proof of Lemma 3.1. Under assumptions (A5′)-(A7′),

$$\sup_{\lambda\in[0,\lambda_{\max}]} \frac{||\boldsymbol{b}||^2}{\tilde{R}_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \leq \frac{M_2\lambda^2_{\max}d}{L_{KL}(\boldsymbol{\beta}^*_{\bar{\alpha}})} \leq \frac{M_1^2 M_2 d}{nL_{KL}(\boldsymbol{\beta}^*_{\bar{\alpha}})} \to 0.$$

□

Lemma B.1. *Under (R1)-(R5), for n sufficiently large*

$$L_{KL}(\hat{\boldsymbol{\beta}}_\alpha) = L_{KL}(\boldsymbol{\beta}^*_\alpha) + \frac{1}{n}||\mathbf{W}^{1/2}_\alpha \mathbf{H}_\alpha(\mathbf{y} - \boldsymbol{\mu})||^2 + O_p(||\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha||^2).$$

Proof. Taylor's expansion of $b(\hat{\boldsymbol{\theta}}_\alpha)$ around $\boldsymbol{\theta}^*_\alpha$ gives us

$$\mathbf{1}^T b(\hat{\boldsymbol{\theta}}_\alpha) = \mathbf{1}^T b(\boldsymbol{\theta}^*_\alpha) + b'(\boldsymbol{\theta}^*_\alpha)^T(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha)$$
$$+ \frac{1}{2}(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha)^T \mathbf{W}_\alpha (\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha) + o_p(||\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha||^2).$$

For n sufficiently large, we have that

$$\begin{aligned}
L_{KL}(\hat{\boldsymbol{\beta}}_\alpha) &= \frac{2}{n}\boldsymbol{\mu}^T(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_\alpha + \frac{2}{n}\mathbf{1}^T b(\hat{\boldsymbol{\theta}}_\alpha) \\
&= L_{KL}(\boldsymbol{\beta}^*_\alpha) - \frac{2}{n}(\boldsymbol{\mu} - b'(\boldsymbol{\theta}^*_\alpha))^T(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha) + \frac{1}{n}(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha)^T\mathbf{W}_\alpha(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}^*_\alpha) \\
&+ o_p(||\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha||^2) \\
&= L_{KL}(\boldsymbol{\beta}^*_\alpha) + \frac{1}{n}||\mathbf{W}_\alpha^{1/2}\mathbf{H}_\alpha(\mathbf{y} - \boldsymbol{\mu})||^2 + O_p(||\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha||^2),
\end{aligned}$$

where the last equality follows from equations (3.1) and (3.2). □

Lemma B.2. *Under assumptions (A1′)-(A4′), (A7) and regularity conditions (R1)-(R3), the following results hold.*

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{(\boldsymbol{y} - \boldsymbol{\mu})^T(\boldsymbol{\theta}^*_\alpha - \boldsymbol{\theta}_0)}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)}\right| \rightarrow_p 0, \tag{B.1}$$

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{(d_\alpha - tr\{(\mathbf{X}'_\alpha\mathbf{W}_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha\mathbf{W}_0\mathbf{X}_\alpha\})}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)}\right| \rightarrow_p 0, \tag{B.2}$$

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{((\mathbf{y} - \boldsymbol{\mu})'\mathbf{H}_\alpha(\mathbf{y} - \boldsymbol{\mu}) - tr\{(\mathbf{X}'_\alpha\mathbf{W}_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha\mathbf{W}_0\mathbf{X}_\alpha\})}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)}\right| \rightarrow_p 0, \tag{B.3}$$

and

$$\sup_{\alpha\in\mathcal{A}_n}\left|\frac{L_{KL}(\hat{\boldsymbol{\beta}}_\alpha)}{R_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} - 1\right| \rightarrow_p 0. \tag{B.4}$$

The proof of this lemma requires the following matrix algebra results.

Definition B.1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $K \times K$ *matrices. We say that* $\mathbf{A} \geq \mathbf{B}$ *if* $\mathbf{A} - \mathbf{B}$ *is positive semidefinite.*

Lemma B.3. *(Horn and Johnson, 1985, p.471) If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $K \times K$ *positive definite Hermitian matrices, then*

(i.) $\mathbf{A} \geq \mathbf{B}$ *if and only if* $\mathbf{B}^{-1} \geq A^{-1}$*;*

(ii.) if $\mathbf{A} \geq \mathbf{B}$*, then* $\lambda_k(\mathbf{A}) \geq \lambda_k(\mathbf{B})$ *for all* $k = 1, \ldots, K$*, where* $\lambda_k(\mathbf{A})$ *and* $\lambda_k(\mathbf{B})$ *are the* k^{th} *largest eigenvalues of* $\mathbf{A}$ *and* $\mathbf{B}$*, respectively.*

Lemma B.4. *(Marshall et al., 2010, p.340) If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $K \times K$ *positive semidefinite Hermitian matrices, then*

$$tr(\mathbf{AB}) \leq \sum_{k=1}^{K} \lambda_k(\mathbf{A})\lambda_k(\mathbf{B}).$$

Proof of Lemma B.2. We start by proving equation (B.1). By Chebyshev's Inequality and Theorem 2 of Whittle (1960), we have that

$$\Pr\left(\sup_{\alpha \in \mathcal{A}_n} \left| \frac{(\boldsymbol{y} - \boldsymbol{\mu})^T(\boldsymbol{\theta}_\alpha^* - \boldsymbol{\theta}_0)}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} \right| > \delta \right) \leq \frac{C}{\delta^{2q}} \sum_{\alpha \in \mathcal{A}_n} \frac{||\boldsymbol{\theta}_\alpha^* - \boldsymbol{\theta}_0||^{2q}}{(nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha))^{2q}}. \tag{B.5}$$

Now $R_{KL}(\hat{\boldsymbol{\beta}}_\alpha) \geq L_{KL}(\boldsymbol{\beta}_\alpha^*)$. If we consider $L_{KL}(\cdot)$ as a function of $\boldsymbol{\theta}$, then by a second order Taylor series expansion around $\boldsymbol{\theta}_0$,

$$L_{KL}(\boldsymbol{\theta}_\alpha^*) = \frac{1}{n}(\boldsymbol{\theta}_\alpha^* - \boldsymbol{\theta}_0)^T \bar{\boldsymbol{W}} (\boldsymbol{\theta}_\alpha^* - \boldsymbol{\theta}_0),$$

where $\bar{\boldsymbol{W}} = diag\{b''(\bar{\theta}_1), \ldots, b''(\bar{\theta}_n)\}$ and $\bar{\theta}_i$ is on the line segment between $\theta_{\alpha i}^*$ and θ_{0i}. Since $b''(\theta) > 0$ for all θ, it follows that $nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha) \geq K||\boldsymbol{\theta}_\alpha^* - \boldsymbol{\theta}_0||^2$ for some constant $K > 0$. Therefore the right-hand side of equation (B.5) is less than or equal to

$$\frac{C'}{\delta^{2q}} \sum_{\alpha \in \mathcal{A}_n} (nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha))^{-q}$$

for some constant $C' > 0$, which tends to zero as $n \to \infty$ by assumption (A3$'$). Next, to establish equation (B.2) we first note that $\mathbf{X}_\alpha'\mathbf{W}_0\mathbf{X}_\alpha \leq \max_{1 \leq i \leq n} \sigma_i^2 \mathbf{X}_\alpha'\mathbf{X}_\alpha$ and $\mathbf{X}_\alpha'\mathbf{W}_\alpha\mathbf{X}_\alpha \geq \min_{1 \leq i \leq n} b''(\theta_{\alpha i})\mathbf{X}_\alpha'\mathbf{X}_\alpha$. From Lemmas B.2 and B.3 it follows then that

$$tr((\mathbf{X}_\alpha'\mathbf{W}_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha'\mathbf{W}_0\mathbf{X}_\alpha) \leq d_\alpha \frac{\max_{1 \leq i \leq n} \sigma_i^2}{\min_{1 \leq i \leq n} b''(\theta_{\alpha i})} \lambda_1\left(\left(\frac{1}{n}\mathbf{X}_\alpha'\mathbf{X}_\alpha\right)^{-1}\right) \lambda_1\left(\frac{1}{n}\mathbf{X}_\alpha'\mathbf{X}_\alpha\right) \leq d_\alpha C$$

for some constant $C > 0$. Using this result we have that

$$\sup_{\alpha \in \mathcal{A}_n} \left| \frac{2(d_\alpha - tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\})}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} \right| \leq \sup_{\alpha \in \mathcal{A}_n} \frac{2d_\alpha(1+C)}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} \leq \frac{d_n(1+C)}{nL_{KL}(\boldsymbol{\beta}^*_{\tilde{\alpha}})},$$

which tends to zero by assumption (A7′).

To prove equation (B.3) we apply Chebyshev's Inequality and Theorem 2 of Whittle (1960) to get that

$$\Pr\left(\sup_{\alpha \in \mathcal{A}_n} \left| \frac{2((\mathbf{y}-\boldsymbol{\mu})'\mathbf{H}_\alpha(\mathbf{y}-\boldsymbol{\mu}) - tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\})}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} \right| > \delta \right)$$
$$\leq \delta^{-2q} C \sum_{\alpha \in \mathcal{A}_n} \frac{tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha (\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\}^q}{(nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha))^{2q}}$$

for some constant $C > 0$. Using the fact that $tr\{\mathbf{AB}\} \leq \lambda_1(\mathbf{A}) tr\{\mathbf{B}\}$,

$$tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha (\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\} \leq K tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\}$$

for some constant $K > 0$. Therefore

$$\Pr\left(\sup_{\alpha \in \mathcal{A}_n} \left| \frac{2((\mathbf{y}-\boldsymbol{\mu})'\mathbf{H}_\alpha(\mathbf{y}-\boldsymbol{\mu}) - tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\})}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} \right| > \delta \right)$$
$$\leq \delta^{-2q} C' \sum_{\alpha \in \mathcal{A}} \frac{tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\}^q}{(nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha))^{2q}}.$$

for some constant $C' > 0$. Since

$$\frac{tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\}}{n} \leq R_{KL}(\hat{\boldsymbol{\beta}}_\alpha),$$

it follows that

$$\Pr\left(\sup_{\alpha \in \mathcal{A}_n} \left| \frac{2((\mathbf{y}-\boldsymbol{\mu})'\mathbf{H}_\alpha(\mathbf{y}-\boldsymbol{\mu})) - tr\{(\mathbf{X}_\alpha' \mathbf{W}_\alpha \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha' \mathbf{W}_0 \mathbf{X}_\alpha\})}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha)} \right| > \delta \right)$$
$$\leq \delta^{-2q} C' \sum_{\alpha \in \mathcal{A}_n} (nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha))^{-q} \to 0.$$

Finally, equation (B.4) follows from (B.3). □

Lemma B.5. *Under (A1′)*

$$||\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}||^2 \leq nC||\mathbf{b}||^2.$$

Proof. $\hat{\boldsymbol{\beta}}_\lambda$ satisfies

$$0 = \frac{1}{n}\frac{\partial l(\hat{\boldsymbol{\beta}}_\lambda)}{\partial \boldsymbol{\beta}} - \mathbf{b}.$$

Without loss of generality, we can write $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}_{\lambda 1}, \hat{\boldsymbol{\beta}}_{\lambda 2})'$ where $\hat{\boldsymbol{\beta}}_{\lambda 2} = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_{\lambda 1}$ is a $1 \times d_{\alpha_\lambda}$ vector of estimated coefficients. Applying the mean value theorem, we get that

$$0 = \frac{1}{n}\frac{\partial l(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})}{\partial \boldsymbol{\beta}} + \frac{1}{n}\frac{\partial^2 l(\bar{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}}_{\lambda 1} - \hat{\boldsymbol{\beta}}_{\alpha_\lambda}) - \mathbf{b}_1,$$

where $\bar{\boldsymbol{\beta}}$ is on the line segment joining $\hat{\boldsymbol{\beta}}_{\lambda 1}$ and $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}$, and $\mathbf{b}_1$ are the non-zero components of $\mathbf{b}$ that correspond to $\hat{\boldsymbol{\beta}}_{\lambda 1}$. For n sufficiently large, it follows then that

$$\hat{\boldsymbol{\beta}}_{\lambda 1} - \hat{\boldsymbol{\beta}}_{\alpha_\lambda} = \left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\bar{\mathbf{W}}_\alpha \mathbf{X}_{\alpha_\lambda}\right)^{-1}\mathbf{b}_1, \tag{B.6}$$

where $\bar{\mathbf{W}}_\alpha = diag\{b''(\bar{\theta}_1), \ldots, b''(\bar{\theta}_n)\}$. Therefore

$$||\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}||^2 = ||X_{\alpha_\lambda}(\hat{\boldsymbol{\beta}}_{\lambda 1} - \hat{\boldsymbol{\beta}}_{\alpha_\lambda})||^2 = n\mathbf{b}_1'\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\bar{\mathbf{W}}_\alpha \mathbf{X}_{\alpha_\lambda}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda}\right)\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\bar{\mathbf{W}}_\alpha \mathbf{X}_{\alpha_\lambda}\right)^{-1}\mathbf{b}_1.$$

Since

$$\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\bar{\mathbf{W}}_\alpha \mathbf{X}_{\alpha_\lambda}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda}\right)\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\bar{\mathbf{W}}_\alpha \mathbf{X}_{\alpha_\lambda}\right)^{-1} \leq (\min_{1 \leq i \leq n} b''(\bar{\theta}_i))^{-2}\left(\frac{1}{n}\mathbf{X}'_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda}\right)^{-1}, \tag{B.7}$$

$$||\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}||^2 \leq nC||\mathbf{b}||^2$$

by Lemma B.3 and assumption (A1′). □

Since $\mathcal{A}_n$ includes all subsets, the results in Lemma B.2 will still hold when the candidate model α is replaced by the random candidate model α_λ.

Lemma B.6. *Under (A1′)-(A7′),*

$$\sup_{\lambda \in [0,\lambda_{\max}]} \left| \frac{L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)}{L_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} - 1 \right| \to_p 0. \tag{B.8}$$

Proof. Applying a second-order Taylor expansion, we get

$$\begin{aligned} L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) - L_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) &= -\frac{2}{n}\boldsymbol{\mu}'(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}) + \frac{2}{n}(b(\hat{\boldsymbol{\theta}}_\lambda) - b(\hat{\boldsymbol{\theta}}_{\alpha_\lambda})) \\ &= -\frac{2}{n}(\boldsymbol{\mu} - b'(\hat{\boldsymbol{\theta}}_{\alpha_\lambda}))'(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}) + \frac{1}{n}(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda})'\bar{\mathbf{W}}_\alpha(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}) \\ &= \frac{2}{n}(\mathbf{y} - \boldsymbol{\mu})'(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}) + \frac{1}{n}(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda})'\bar{\mathbf{W}}_\alpha(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}), \end{aligned}$$

where the last equality follows from the fact that $\hat{\boldsymbol{\theta}}_{\alpha_\lambda}$ is the maximum-likelihood estimator so $\mathbf{X}'_{\alpha_\lambda}(\mathbf{y} - b'(\hat{\boldsymbol{\theta}}_{\alpha_\lambda})) = 0$.

By equation (B.6) and assumptions (A5′) and (A6′), the first term is bounded by

$$M_1 \frac{2}{n}(\boldsymbol{y} - \boldsymbol{\mu})^T \frac{\boldsymbol{X}_{\alpha_\lambda}}{\sqrt{n}}(\frac{1}{n}\boldsymbol{X}^T_{\alpha_\lambda}\bar{\boldsymbol{W}}_\alpha \boldsymbol{X}^T_{\alpha_\lambda})^{-1}\mathbf{1}$$

where $\mathbf{1}$ is a $d_{\alpha_\lambda} \times 1$ vector of ones. Applying Chebyshev's Inequality and Theorem 2 of Whittle (1960), we have that

$$\Pr\left(\sup_{\lambda \in [0,\lambda_{\max}]} \frac{(\boldsymbol{y} - \boldsymbol{\mu})^T(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda})}{nR_{KL}(\boldsymbol{\beta}_{\alpha_\lambda})} > \delta \right) \leq \frac{C}{\delta^{2q}} \sum_{\alpha \in \mathcal{A}_n} \frac{||n^{-1/2}\boldsymbol{X}_\alpha(\frac{1}{n}\boldsymbol{X}^T_\alpha \bar{\boldsymbol{W}}_\alpha \boldsymbol{X}^T_\alpha)^{-1}\mathbf{1}||^{2q}}{n^{2q}R_{KL}(\hat{\boldsymbol{\beta}}_\alpha)^{2q}}$$

for some constant $C > 0$. By equation (B.7) and assumption (A1′), this does not exceed

$$\frac{C'}{\delta^{2q}} \sum_{\alpha \in \mathcal{A}_n} \frac{d^q_\alpha}{n^{2q}R_{KL}(\hat{\beta}_\alpha)^{2q}}.$$

By (A6′), $d/nR_{KL}(\beta^*_{\bar{\alpha}}) \to 0$, so, for n sufficiently large, $d_\alpha < nR_{KL}(\hat{\beta}_\alpha)$. Therefore, the last quantity is less than or equal to

$$\frac{C'}{\delta^{2q}} \sum_{\alpha \in \mathcal{A}_n} (nR_{KL}(\hat{\beta}_\alpha))^{-q},$$

which tends to zero by assumption (A3′). Thus

$$\sup_{\lambda \in [0,\lambda_{\max}]} \frac{2(\mathbf{y}-\boldsymbol{\mu})^T(\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}_{\alpha_\lambda})}{nL_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda})} \to_p 0.$$

Assuming that (A4′)-(A7′) holds, equation (B.8) follows from this result and Lemma B.5. □

Proof. To prove the efficiency of AIC_λ, it suffices to show that

$$\sup_{\lambda \in [0,\lambda_{\max}]} \left| \frac{AIC_\lambda - \frac{2}{n}\boldsymbol{y}^T\boldsymbol{\theta}_0 + \frac{2}{n}\mathbf{1}^T b(\theta_0) - L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)}{L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)} \right| \to_p 0. \tag{B.9}$$

Consider

$$\begin{aligned} AIC_\lambda - \frac{2}{n}\boldsymbol{y}^T\boldsymbol{\theta}_0 + \frac{2}{n}\mathbf{1}^T b(\boldsymbol{\theta}_0) &= \frac{2}{n}\boldsymbol{y}^T(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_\lambda) + \frac{2}{n}\mathbf{1}^T(b(\hat{\boldsymbol{\theta}}_\lambda) - b(\boldsymbol{\theta}_0)) + 2\frac{d_{\alpha_\lambda}}{n} \\ &= L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) + \frac{2}{n}(\boldsymbol{y}-\boldsymbol{\mu})^T(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*_{\alpha_\lambda}) \\ &+ \frac{2}{n}(\boldsymbol{y}-\boldsymbol{\mu})^T(\boldsymbol{\theta}^*_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}) + \frac{2}{n}(\boldsymbol{y}-\boldsymbol{\mu})^T(\hat{\boldsymbol{\theta}}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}_\lambda) + 2\frac{d_{\alpha_\lambda}}{n}. \end{aligned}$$

By the expansion in equation (3.2) we have that

$$\hat{\boldsymbol{\theta}}_{\alpha_\lambda} = \boldsymbol{\theta}^*_{\alpha_\lambda} + \mathbf{H}_{\alpha_\lambda}(\mathbf{y} - b'(\boldsymbol{\theta}^*_{\alpha_\lambda}))$$

asymptotically. Therefore

$$\begin{aligned} AIC_\lambda - \frac{2}{n}\boldsymbol{y}^T\boldsymbol{\theta}_0 + \frac{2}{n}\mathbf{1}^T b(\boldsymbol{\theta}_0) &= L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) + \frac{2}{n}(\boldsymbol{y}-\boldsymbol{\mu})^T(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*_{\alpha_\lambda}) \\ &- \frac{2}{n}((\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{H}_{\alpha_\lambda}(\boldsymbol{y}-\boldsymbol{\mu}) - tr\{(\boldsymbol{X}'_{\alpha_\lambda}\boldsymbol{W}_{\alpha_\lambda}\boldsymbol{X}_{\alpha_\lambda})^{-1}\boldsymbol{X}'_{\alpha_\lambda}\boldsymbol{W}_0\boldsymbol{X}_{\alpha_\lambda}\}) \\ &+ \frac{2}{n}(d_{\alpha_\lambda} - tr\{(\boldsymbol{X}'_{\alpha_\lambda}\boldsymbol{W}_{\alpha_\lambda}\boldsymbol{X}_{\alpha_\lambda})^{-1}\boldsymbol{X}'_{\alpha_\lambda}\boldsymbol{W}_0\boldsymbol{X}_{\alpha_\lambda}\}) + \frac{2}{n}(\boldsymbol{y}-\boldsymbol{\mu})'(\hat{\boldsymbol{\theta}}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}_\lambda). \end{aligned}$$

Applying Lemmas B.2 and B.6, equation (B.9) holds as desired. □

C Supplementary Material

This supplemental section contains the technical details required to show that Theorem 3 can be used to prove the efficiency of AIC_λ, GCV_λ, and AIC_{c_λ}, the regularity conditions required for Theorem 4 to hold, and the mathematical results needed to apply Lemma 2.1 to the simulation examples.

C.1 Verifying the Conditions of Theorem 3

The following shows that AIC_λ, GCV_λ, and AIC_{c_λ} can be written in the form $\tilde{\Gamma}_n(\lambda)$ and that Conditions (C1) and (C2) of Theorem 3 are satisfied. This implies that the three methods are efficient selectors of the regularization parameter. Shibata (1981) and Hurvich and Tsai (1989) noted that AIC and AIC_c, respectively, can be shown to satisfy these conditions. We present a detailed argument of these remarks below.

AIC_λ is Efficient

Minimizing AIC_λ is equivalent to minimizing

$$\exp\left(\frac{2d_{\alpha_\lambda}}{n}\right)\hat{\sigma}^2_\lambda.$$

Using Taylor's expansion we get

$$\begin{aligned}\exp\left(\frac{2d_{\alpha_\lambda}}{n}\right)\hat{\sigma}^2_\lambda &= \sum_{k=0}^{\infty}\left(\frac{2d_{\alpha_\lambda}}{n}\right)^k\frac{1}{k!}\\ &= 1+\frac{2d_{\alpha_\lambda}}{n}+\sum_{k=2}^{\infty}\left(\frac{2d_{\alpha_\lambda}}{n}\right)^k\frac{1}{k!},\end{aligned}$$

and we see that AIC_λ has the same asymptotic properties as

$$\tilde{\Gamma}_\lambda = \hat{\sigma}_\lambda^2 \left(1 + 2\frac{d_{\alpha_\lambda}}{n} + \delta_\lambda\right),$$

where

$$\delta_n(\lambda) = \sum_{k=2}^{\infty} \left(\frac{2d_{\alpha_\lambda}}{n}\right)^k \frac{1}{k!}.$$

Therefore, the efficiency of AIC_λ can be established by showing that (C1) and (C2) hold. Consider

$$0 < \delta_\lambda = \sum_{k=2}^{\infty} \left(\frac{2d_{\alpha_\lambda}}{n}\right)^k \frac{1}{k!} = \exp\left(\frac{2d_{\alpha_\lambda}}{n}\right) - 1 - \frac{2d_{\alpha_\lambda}}{n}.$$

Therefore, under the assumption that $d_n/n \to 0$, (C1) is satisfied. Next consider

$$\begin{aligned} 0 < \frac{\delta_\lambda}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} &= \sum_{k=2}^{\infty} \left(\frac{2d_{\alpha_\lambda}}{n}\right)^k \frac{1}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})k!} \\ &\leq \frac{2}{\sigma^2} \sum_{k=2}^{\infty} \left(\frac{2d_{\alpha_\lambda}}{n}\right)^{k-1} \frac{1}{k!} \leq \frac{2}{\sigma^2} \sum_{k=2}^{\infty} \left(\frac{2d_n}{n}\right)^{k-1} \frac{1}{(k-1)!} \\ &= \frac{2}{\sigma^2} \sum_{k=1}^{\infty} \left(\frac{2d_n}{n}\right)^k \frac{1}{k!} = \frac{2}{\sigma^2}\left(\exp\left(\frac{2d_n}{n}\right) - 1\right) \to 0. \end{aligned}$$

Here the inequality on the second line follows from the fact that $R(\hat{\beta}_{\alpha_\lambda}) > \sigma^2 d_{\alpha_\lambda}/n$ and the final result follows from the assumption that $d_n/n \to 0$. Therefore,

$$\sup_{\lambda\in[0,\lambda_{max}]} \frac{|\delta_\lambda|}{L(\hat{\beta}_\lambda)} = \sup_{\lambda\in[0,\lambda_{max}]} \left| \frac{L(\hat{\beta}_{\alpha_\lambda})}{L(\hat{\beta}_\lambda)} \frac{\tilde{R}(\hat{\beta}_{\alpha_\lambda})}{L(\hat{\beta}_{\alpha_\lambda})} \frac{\delta_\lambda}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} \right| \to_p 0$$

so (C2) is satisfied.

GCV_λ is Efficient

Using Taylor's expansion we get

$$\frac{1}{(1-d_{\alpha_\lambda}/n)^2} = \sum_{k=1}^{\infty} k\left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-1} = 1 + \frac{2d_{\alpha_\lambda}}{n} + \sum_{k=3}^{\infty} k\left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-1},$$

and we see that GCV_λ has the same asymptotic properties as

$$\tilde{\Gamma}_\lambda = \hat{\sigma}^2_\lambda \left(1 + 2\frac{d_{\alpha_\lambda}}{n} + \delta_\lambda\right),$$

where

$$\delta_\lambda = \sum_{k=3}^{\infty} k \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-1}.$$

Therefore, the efficiency of GCV_λ can be established by showing that (C1) and (C2) hold. Consider

$$0 < \delta_\lambda = \sum_{k=3}^{\infty} k \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-1} = \frac{1}{(1 - d_{\alpha_\lambda}/n)^2} - 1 - \frac{2d_{\alpha_\lambda}}{n}.$$

Therefore, under the assumption that $d_n/n \to 0$, (C1) is satisfied. Next consider

$$\begin{aligned} 0 < \frac{\delta_\lambda}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} &= \sum_{k=3}^{\infty} k \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-1} \frac{1}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} \\ &\leq \frac{1}{\sigma^2} \sum_{k=3}^{\infty} k \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-2} \\ &= \frac{1}{\sigma^2} \left(\sum_{k=3}^{\infty} (k-1) \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-2} + \sum_{k=3}^{\infty} \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-2} \right) \\ &= \frac{1}{\sigma^2} \left(\sum_{k=2}^{\infty} k \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k-1} + \frac{d_{\alpha_\lambda}}{n} \sum_{k=0}^{\infty} \left(\frac{d_{\alpha_\lambda}}{n}\right)^{k} \right) \\ &= \frac{1}{\sigma^2} \left(\frac{1}{(1 - d_{\alpha_\lambda}/n)^2} - 1 + \frac{d_{\alpha_\lambda}/n}{1 - d_{\alpha_\lambda}/n} \right), \end{aligned}$$

which converges to zero uniformly over λ under the assumption that $d_n/n \to 0$. Here, again, the inequality on the second line follows from the fact that $\tilde{R}(\hat{\beta}_{\alpha_\lambda}) > \sigma^2 d_{\alpha_\lambda}/n$. Therefore,

$$\sup_{\lambda \in [0, \lambda_{max}]} \frac{|\delta_\lambda|}{L(\hat{\beta}_\lambda)} = \sup_{\lambda \in [0, \lambda_{max}]} \left| \frac{L(\hat{\beta}_{\alpha_\lambda})}{L(\hat{\beta}_\lambda)} \frac{\tilde{R}(\hat{\beta}_{\alpha_\lambda})}{L(\hat{\beta}_{\alpha_\lambda})} \frac{\delta_\lambda}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} \right| \to_p 0$$

so (C2) is satisfied.

AIC_{c_λ} is Efficient

We define

$$AIC_{c_\lambda} = \log(\hat{\sigma}^2_\lambda) + 2\frac{d_{\alpha_\lambda}+1}{n-d_{\alpha_\lambda}-2}.$$

This can be equivalently defined as

$$AIC_{c_\lambda} = \log(\hat{\sigma}^2_\lambda) + 2\frac{d_{\alpha_\lambda}+1}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}.$$

Based on the second definition of AIC_{c_λ} we see that the information criterion has the same asymptotic properties as

$$\log(\hat{\sigma}^2_\lambda) + 2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)},$$

because they only differ by an additive constant $(2/n)$. Therefore, AIC_{c_λ} will have the same asymptotic behavior as

$$\exp\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)\hat{\sigma}^2_\lambda.$$

Using Taylor's expansion we get

$$\begin{aligned}\exp\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right) &= \sum_{k=0}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^k\frac{1}{k!}\\ &= 1 + \frac{2d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\\ &+ \sum_{k=2}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^k\frac{1}{k!},\end{aligned}$$

and we see that AIC_{c_λ} has the same asymptotic properties as

$$\tilde{\Gamma}_\lambda = \hat{\sigma}^2_\lambda\left(1 + 2\frac{d_{\alpha_\lambda}}{n} + \delta_\lambda\right),$$

where

$$\delta_\lambda = 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)} + \sum_{k=2}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^k\frac{1}{k!}.$$

Therefore, the efficiency of AIC_{c_λ} can be established by showing that (C1) and (C2) hold. Consider

$$\begin{aligned}
0 < \delta_n(\lambda) &= 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)} + \sum_{k=2}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^k \frac{1}{k!} \\
&= 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)} + \exp\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right) \\
&\quad - 1 - 2\frac{d_{\alpha_\lambda}}{n} - 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)},
\end{aligned}$$

which converges to zero uniformly over λ under the assumption that $d_n/n \to 0$. Thus, (C1) is satisfied. Next consider

$$\begin{aligned}
0 < \frac{\delta_\lambda}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} &= 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})n(n-d_{\alpha_\lambda}-2)} \\
&\quad + \sum_{k=2}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^k \frac{1}{\tilde{R}(\hat{\beta}^*_{\alpha_\lambda})k!} \\
&\le 2\frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{\sigma^2(n-d_{\alpha_\lambda}-2)} \\
&\quad + \frac{n}{\sigma^2 d_{\alpha_\lambda}}\sum_{k=2}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^k \frac{1}{k!} \\
&\le 2\frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{\sigma^2(n-d_{\alpha_\lambda}-2)} \\
&\quad + \frac{2}{\sigma^2}\left(1 + \frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{(n-d_{\alpha_\lambda}-2)}\right)\sum_{k=2}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^{k-1} \frac{1}{k!} \\
&\le 2\frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{\sigma^2(n-d_{\alpha_\lambda}-2)} \\
&\quad + \frac{2}{\sigma^2}\left(1 + \frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{(n-d_{\alpha_\lambda}-2)}\right)\sum_{k=1}^{\infty}\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right)^{k} \frac{1}{k!} \\
&= 2\frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{\sigma^2(n-d_{\alpha_\lambda}-2)} \\
&\quad + \frac{2}{\sigma^2}\left(1 + \frac{(1+1/d_{\alpha_\lambda})(d_{\alpha_\lambda}+2)}{(n-d_{\alpha_\lambda}-2)}\right)\left(\exp\left(2\frac{d_{\alpha_\lambda}}{n} + 2\frac{(d_{\alpha_\lambda}+1)(d_{\alpha_\lambda}+2)}{n(n-d_{\alpha_\lambda}-2)}\right) - 1\right),
\end{aligned}$$

which converges to zero uniformly over λ under the assumption that $d_n/n \to 0$. Again, the inequality on the third line follows from the fact that $R(\hat{\beta}_n^*(\alpha_\lambda)) > \sigma^2 d_{\alpha_\lambda}/n$. Therefore,

$$\sup_{\lambda\in[0,\lambda_{max}]} \frac{|\delta_\lambda|}{L(\hat{\beta}_\lambda)} = \sup_{\lambda\in[0,\lambda_{max}]} \left| \frac{L(\hat{\beta}_{\alpha_\lambda})}{L(\hat{\beta}_\lambda)} \frac{\tilde{R}(\hat{\beta}_{\alpha_\lambda})}{L(\hat{\beta}_{\alpha_\lambda})} \frac{\delta_\lambda}{\tilde{R}(\hat{\beta}_{\alpha_\lambda})} \right| \to_p 0$$

so (C2) is satisfied.

C.2 Regularity Conditions

Below are the regularity conditions required to derive the properties of the maximum-likelihood estimator for misspecified models. Refer to Lv and Liu (2010) for a discussion of these conditions in the context of generalized linear models with no dispersion parameter.

(R1) $f_\alpha(y;\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$ for every $\boldsymbol{\beta}$ in Ω, a compact set of $\mathbb{R}^{d_\alpha}$.

(R2) (a.) $E_0(\log(g(y)))$ exists and $|\log f_\alpha(y;\boldsymbol{\beta})|$ is dominated by an integrable function with respect to g that is independent of $\boldsymbol{\beta}$. (b.) The KL loss function has a unique minimum at $\boldsymbol{\beta}^*$, which is an interior point of Ω.

(R3) (a.) $\partial \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i$ and $\partial^2 \log f(y;\boldsymbol{\beta})/\partial\beta_i\partial\beta_j$, $i,j = 1,\ldots,d_\alpha$, are measurable functions of y for each $\boldsymbol{\beta}\in\Omega$ and continuously differentiable functions of $\boldsymbol{\beta}$ for each y. (b.) $|\partial \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i|$, $|\partial \log f(y;\boldsymbol{\beta})/\partial\beta_i\partial\beta_j|$, and $|(\partial \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i)(\partial \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_j)|$ are dominated by integrable functions with respect to g, which are independent of $\boldsymbol{\beta}$.

(R4) The matrices

$$B(\theta^*) = E_0\left(\frac{\partial \log f_\alpha(y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\frac{\partial \log f_\alpha(y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right)$$

and

$$A(\theta^*) = E_0\left(\frac{\partial^2 \log f_\alpha(y;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)$$

are positive definite.

(R5) (a.) $\partial^3 \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i\beta_j\beta_k$ are measurable with respect to y for $i,j,k = 1,\ldots,d_{alpha}$.

(b.) $|\partial \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i|^2$, $|\partial^2 \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i\partial\beta_j|^2$, and $|\partial^3 \log f_\alpha(y;\boldsymbol{\beta})/\partial\beta_i\partial\beta_j\partial\beta_k|^2$, $i,j,k = 1,\ldots,d_\alpha$, are dominated by integrable functions with respect to g that are independent of $\boldsymbol{\beta}$.

(R6) For some $\delta > 0$, $E||\mathbf{B}_n^{-1/2}\mathbf{A}_n(\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}^*_\alpha)||^{3+\delta} = O(1)$, where $\mathbf{A}_n$ is defined as in equation (3) of the manuscript and $\mathbf{B}_n = \mathbf{X}_\alpha^T W_0 \mathbf{X}_\alpha$.

C.3 Verifying the Conditions of Lemma 2.1

C.3.1 Omitted Predictor with Deterministic $\boldsymbol{X}$

We first consider a more general example. Let the true model be defined as

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the $n \times 1$ response vector, $\boldsymbol{\mu}$ is the $n \times 1$ unknown mean vector, and $\boldsymbol{\varepsilon}$ is a $n \times 1$ noise vector where $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$. In what follows we assume that

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \beta_{\text{excl}}\boldsymbol{x}_{\text{excl}},$$

where $\mathbf{X}$ is a $n \times d_n$ deterministic matrix of predictors, $\boldsymbol{\beta}$ is a $d_n \times 1$ vector of coefficients, $\boldsymbol{x}_{\text{excl}}$ is a $n \times 1$ deterministic vector, and β_{excl} is a constant. In the following, we take the candidate models to be the least squares regressions based on all 2^{d_n} subsets of $\boldsymbol{X}$; the predictor $\boldsymbol{x}_{\text{excl}}$ is excluded from consideration so that the true model is never included in the set of candidate models.

Assume that the following conditions hold:

(C3) $\boldsymbol{\beta}$ contains a fixed number of non-zero entries

(C4) $\boldsymbol{x}_{\text{excl}}$ is orthogonal to the columns of $\boldsymbol{X}$

(C5) $\inf_n \frac{\boldsymbol{x}_{\text{excl}}^T \boldsymbol{x}_{\text{excl}}}{n} > 0$

By construction, for any candidate model α,

$$\begin{aligned}
nR(\hat{\beta}_\alpha) &\geq ||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}\boldsymbol{\mu}||^2 \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_{\bar{\alpha}})X\boldsymbol{\beta}||^2 + \boldsymbol{\beta}^T X^T (\boldsymbol{I} - \boldsymbol{H}_{\bar{\alpha}})\boldsymbol{x}_{\text{excl}}\beta_{\text{excl}} + ||\boldsymbol{x}_{\text{excl}}\beta_{\text{excl}}||^2 \\
&= ||(\boldsymbol{I} - \boldsymbol{H}_{\bar{\alpha}})X\boldsymbol{\beta}||^2 + ||\boldsymbol{x}_{\text{excl}}\beta_{\text{excl}}||^2 \\
&\geq ||\boldsymbol{x}_{\text{excl}}\beta_{\text{excl}}||^2 \\
&= n\beta_{\text{excl}}^2 \left(\frac{\boldsymbol{x}_{\text{excl}}^T \boldsymbol{x}_{\text{excl}}}{n} \right) \\
&\geq k_1 n
\end{aligned}$$

for some constant $k_1 > 0$.

For the simulation example in Section 4.1 of the paper, the true vector of coefficients is fixed and trigonometric predictors are used so conditions (C3)-(C5) are satisfied. Therefore, for that example it follows that $||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}\boldsymbol{\mu}||^2 \geq k_1 n$ for some constant $k_1 > 0$.

C.3.2 Exponential Model

From Fourier analysis (cf. Bloomfield (2000)), if n is even then

$$\mu_t = e^{4t/n} = A(0) + \sum_{0<j<n/2} A(f_j)\cos(2\pi f_j t) + \sum_{0<j<n/2} B(f_j)\sin(2\pi f_j t) + A(f_{n/2})\cos\left(2\pi f_{n/2} t\right), \tag{C.1}$$

where $f_j = j/n$,

$$A(f_j) = \frac{2}{n}\sum_{t=0}^{n-1} \mu_t \cos(2\pi f_j t),$$

and

$$B(f_j) = \frac{2}{n}\sum_{t=0}^{n-1} \mu_t \sin(2\pi f_j t).$$

If n is odd then the rightmost term in (C.1) is excluded. To determine $A(f_j)$ and $B(f_j)$ we will use the fact that $d(f_j) = \frac{A(f_j)}{2} - i\frac{B(f_j)}{2}$, where

$$d(f_j) = \frac{1}{n}\sum_{t=0}^{n-1} \mu_t e^{-2\pi i f_j t}.$$

For this example

$$d(f_j) = \frac{1}{n}\sum_{t=0}^{n-1} e^{4t/n} e^{-2\pi i f_j t} = \frac{1 - e^{4-2\pi i}}{n(1 - e^{4/n - 2\pi i f_j})} = \frac{1 - e^4}{n} \frac{1}{(1 - e^{4/n}\cos(2\pi f_j)) + i e^{4/n}\sin(2\pi f_j)}.$$

For any real constants a and b, $\frac{1}{a+bi} = \frac{a-bi}{a^2+b^2}$. It follows then that

$$\begin{aligned} d(f_j) &= \frac{1 - e^4}{n} \frac{1 - e^{4/n}\cos(2\pi f_j) - i e^{4/n}\sin(2\pi f_j)}{(1 - e^{4/n}\cos(2\pi f_j))^2 + (e^{4/n}\sin(2\pi f_j))^2} \\ &= \frac{1 - e^4}{n}\left(\frac{1 - e^{4/n}\cos(2\pi f_j)}{1 + (e^{4/n})^2 - 2e^{4/n}\cos(2\pi f_j)} - i\frac{e^{4/n}\sin(2\pi f_j)}{1 + (e^{4/n})^2 - 2e^{4/n}\cos(2\pi f_j)}\right). \end{aligned}$$

Therefore

$$A(f_j) = 2\frac{1 - e^4}{n} \frac{(e^{-4/n} - \cos(2\pi f_j))}{e^{-4/n} + e^{4/n} - 2\cos(2\pi f_j)}$$

and

$$B(f_j) = 2\frac{1 - e^4}{n} \frac{\sin(2\pi f_j)}{e^{-4/n} + e^{4/n} - 2\cos(2\pi f_j)}.$$

For a given d_n, define the $n \times (n - d_n)$ matrix $\boldsymbol{X}_{\text{excl}} = (\boldsymbol{x}^1_{\text{excl}}, \boldsymbol{x}^2_{\text{excl}})$ with components

$$x^1_{\text{excl}_{tj}} = \sin(2\pi t f_j)$$

and

$$x^2_{\text{excl}_{tj}} = \cos(2\pi t f_j)$$

for $j = d_n/2 + 1, \ldots, n$. Based on this notation, the $n \times 1$ mean vector $\boldsymbol{\mu}$ can be written as

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}_{excl}\boldsymbol{\beta}_{excl},$$

where

$$\boldsymbol{\beta} = [A(0)\ A(f_1) \cdots A(f_{d_n/2})\ B(f_1) \cdots B(f_{d_n/2})]^T$$

and

$$\boldsymbol{\beta}_{excl} = [A(f_{d_n/2+1}) \cdots A(f_{n/2})\ B(f_{d_n/2+1}) \cdots B(f_{n/2-1})]^T.$$

For this example, consider

$$\begin{aligned}
nR(\hat{\beta}_\alpha) &\geq ||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}\boldsymbol{\mu}||^2 \\
&\geq ||\boldsymbol{X}_{excl}\boldsymbol{\beta}_{excl}||^2 \\
&\geq \frac{n}{2}\boldsymbol{\beta}_{excl}^T\boldsymbol{\beta}_{excl} \\
&= \frac{n}{2}\left(\sum_{d_n/2<j<n/2} A(f_j)^2 + B(f_j)^2\right) + \frac{n}{2}A(f_{n/2})^2 \\
&\geq \frac{n}{2}B(f_{d_n/2+1})^2 \\
&= \frac{n}{2}\frac{(2(1-e^4))^2}{n^2}\left(\frac{\sin(2\pi f_{d_n/2+1})}{e^{-4/n} + e^{4/n} - 2\cos(2\pi f_{d_n/2+1})}\right)^2 \\
&\geq n\frac{c_1}{n^2}\left(\frac{\sin(2\pi f_{d_n/2+1})}{2(cosh(4/n) - 1) + 2(1 - \cos(2\pi f_{d_n/2+1})}\right)^2
\end{aligned}$$

for some positive constant c_1. To simplify notation, define

$$h_n = \frac{c_1}{n^2}\left(\frac{\sin(2\pi f_{d_n/2+1})}{2(cosh(4/n) - 2) + 2(1 - \cos(2\pi f_{d_n/2+1})}\right)^2.$$

If $d_n \to \infty$, then $\lim_{n\to\infty} h_n/d_n^2 < \infty$. It follows that

$$||\boldsymbol{\mu} - \boldsymbol{H}_{\bar{\alpha}}\boldsymbol{\mu}||^2 \geq k_1 n d_n^{-2}$$

for some constant $k_1 > 0$.

Acknowledgements

We would like to thank the Associate Editor and three anonymous referees for comments that helped improve the quality and content of our manuscript greatly.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR*, pages 267–281.

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction.* Wiley-Series in Probability and Statistics, 2 edition.

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in Statistics: Proceedings of a Workshop.* Academic Press Inc.,U.S.

Breheny, P. and Huang, J. (2011). Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Applied Statistics*, 5(1):232–253.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

Craven, P. and Wahba, G. (1978). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*, 31(4):377–403.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, 32(3):928–961.

Furnival, G. M. and Wilson, R. W. (1974). Regression by Leaps and Bounds. *Technometrics*, 16(4):499–511.

Gelman, A. (2010). Bayesian Statistics Then and Now. *Statistical Science*, 25(2):162–165.

Hastie, T. and Efron, B. (2011). *lars: Least Angle Regression, Lasso and Forward Stagewise.* R package version 0.9-8.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediciton.* Springer Series in Statistics. Springer, 2 edition.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis.* Cambridge University Press.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2):297–307.

Hurvich, C. M. and Tsai, C.-L. (1991). Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models. *Biometrika*, 78(3):499–509.

Hurvich, C. M. and Tsai, C.-L. (1995). Model Selection for Extended Quasi-Likelihood Models in Small Samples. *Biometrics*, 51(3):1077–84.

Leng, C., Lin, Y., and Wahba, G. (2006). A Note on the Lasso and Related Procedures in Model Selection. *Statistica Sinica*, 16:1273–1284.

Li, K.-C. (1987). Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975.

Lv, J. and Liu, J. (2010). Model Selection Principles in Misspecified Models. arXiv:1005.5483v1.

Mallows, C. L. (1973). Some Comments on C_p. *Technometrics*, 15(4):661–675.

Marshall, A. W., Olkin, I., and Arnold, B. (2010). *Inequalities: Theory of Majorization and Its Applications.* Springer Series in Statistics. Springer, 2 edition.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2 edition.

Nishii, R. (1988). Maximum Likelihood Principle and Model Selection when the True Model Is Unspecified. *Journal of Multivariate Analysis*, 27:392–403.

Park, M. Y. and Hastie, T. (2011). *glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model*. R package version 0.95.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Sela, R. J. and Simonoff, J. S. (2012). RE-EM Trees: A Data Mining Approach for Longitudinal and Clustered Data. *Machine Learning*, 86(2):169–207.

Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, 7:221–264.

Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. *The Annals of Statistics*, 8(1):147–164.

Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika*, 68(1):45–54.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, 94(3):553–568.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25.

Whittle, P. (1960). Bounds for the Moments of Linear and Quadratic Forms in Independent Variables. *Theory of Probability and Its Applications*, 5(3):302–305.

Yang, Y. (2005). Can the Strengths of AIC and BIC be Shared? A Conflict Between Model Indentification and Regression Estimation. *Biometrika*, 92(4):937950.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization Parameter Selections via Generalized Information Criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society B*, 67:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “Degrees of Freedom” of the Lasso. *The Annals of Statistics*, 35(5):2173–2192.